Non-Concave Penalized Likelihood with NP-Dimensionality

Jianqing Fan and Jinchi Lv *
Princeton University and University of Southern California

September 1, 2009

Abstract

Penalized likelihood methods are fundamental to ultra-high dimensional variable selection. How high dimensionality such methods can handle remains largely unknown. In this paper, we show that in the context of generalized linear models, such methods possess model selection consistency with oracle properties even for dimensionality of Non-Polynomial (NP) order of sample size, for a class of penalized likelihood approaches using folded-concave penalty functions, which were introduced to ameliorate the bias problems of convex penalty functions. This fills a long-standing gap in the literature where the dimensionality is allowed to grow slowly with the sample size. Our results are also applicable to penalized likelihood with the L_1-penalty, which is a convex function at the boundary of the class of folded-concave penalty functions under consideration. The coordinate optimization is implemented for finding the solution paths, whose performance is evaluated by a few simulation examples and the real data analysis.

Running title: Non-Concave Penalized Likelihood

Key words: Variable selection; High dimensionality; Non-concave penalized likelihood; Folded-concave penalty; Oracle property; Weak oracle property; Lasso; SCAD

1 Introduction

The analysis of data sets with the number of variables p comparable to or much larger than the sample size n frequently arises nowadays in many fields ranging from genomics and

*Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: jqfan@princeton.edu). Jinchi Lv is Assistant Professor of Statistics, Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA (e-mail: jinchilv@marshall.usc.edu). Fan's research was partially supported by NSF Grants DMS-0704337 and DMS-0714554 and NIH Grant R01-GM072611. Lv's research was partially supported by NSF Grant DMS-0806030 and 2008 Zumberge Individual Award from USC's James H. Zumberge Faculty Research and Innovation Fund.

health sciences to economics and machine learning. The data that we collect is usually of the type $(y_i, x_{i1}, \cdots, x_{ip})_{i=1}^n$, where the y_i's are n independent observations of the response variable Y given its covariates, or explanatory variables, $(x_{i1}, \cdots, x_{ip})^T$. Generalized linear models (GLMs) provide a flexible parametric approach to estimating the covariate effects (McCullagh and Nelder, 1989). In this paper we consider the variable selection problem of Non-Polynomial (NP) dimensionality in the context of GLMs. By NP-dimensionality we mean that $\log p = O(n^a)$ for some $a \in (0, 1)$. See Fan and Lv (2009) for an overview of recent developments in high dimensional variable selection.

We denote by $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ the $n \times p$ design matrix with $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$ and $\mathbf{y} = (y_1, \cdots, y_n)^T$ the n-dimensional response vector. Throughout the paper we consider deterministic design matrix. With a canonical link, the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$ belongs to the canonical exponential family, having the following density function with respect to some fixed measure

$$f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^n f_0(y_i; \theta_i) = \prod_{i=1}^n \left\{ c(y_i) \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is an unknown p-dimensional vector of regression coefficients, $\{f_0(y; \theta) : \theta \in \mathbf{R}\}$ is a family of distributions in the regular exponential family with dispersion parameter $\phi \in (0, \infty)$, and $(\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$. As is common in GLM, the function $b(\theta)$ is implicitly assumed to be twice continuously differentiable with $b''(\theta)$ always positive. In the sparse modeling, we assume that majority of the true regression coefficients $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ are exactly zero. Without loss of generality, assume that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ with each component of $\boldsymbol{\beta}_1$ nonzero and $\boldsymbol{\beta}_2 = \mathbf{0}$. Hereafter we refer to the support $\text{supp}(\boldsymbol{\beta}_0) = \{1, \cdots, s\}$ as the true underlying sparse model of the indices. Variable selection aims at locating those predictors $\mathbf{x}_j$ with nonzero $\beta_{0,j}$ and giving an effective estimate of $\boldsymbol{\beta}_1$.

In view of (1), the log-likelihood $\log f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$ of the sample is given, up to an affine transformation, by

$$\ell_n(\boldsymbol{\beta}) = n^{-1} \left[\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta}) \right], \tag{2}$$

where $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \cdots, b(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$. We consider the following penalized likelihood

$$Q_n(\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) - \sum_{j=1}^p p_{\lambda_n}(|\beta_j|), \tag{3}$$

where $p_\lambda(\cdot)$ is a penalty function and $\lambda_n \geq 0$ is a regularization parameter.

In a pioneering paper, Fan and Li (2001) build the theoretical foundation of nonconcave penalized likelihood for variable selection. The penalty functions that they used are not any nonconvex functions, but really the folded-concave functions. For this reason, we will call them more precisely folded-concave penalties. The paper also introduces the oracle property

for model selection. An estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ is said to have the oracle property (Fan and Li, 2001) if it enjoys the model selection consistency in the sense of $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$, and it attains an information bound mimicking that of the oracle estimator, where $\widehat{\boldsymbol{\beta}}_1$ is a subvector of $\widehat{\boldsymbol{\beta}}$ formed by its first s components and the oracle knew the true model $\text{supp}(\boldsymbol{\beta}_0) = \{1, \cdots, s\}$ ahead of time. Fan and Li (2001) study the oracle properties of non-concave penalized likelihood estimators in the finite-dimensional setting. Their results were extended later by Fan and Peng (2004) to the setting of $p = o(n^{1/5})$ or $o(n^{1/3})$ in a general likelihood framework. The question of how large p can be so that similar oracle properties continue to hold arises naturally. Can the penalized likelihood methods be applicable to NP-dimensional variable selection problems? This paper gives an affirmative answer.

Numerous efforts have lately been devoted to studying the properties of variable selection with ultra-high dimensionality and significant progress has been made. Meinshausen and Bühlmann (2006), Zhao and Yu (2006), and Zhang and Huang (2008) investigate the issue of model selection consistency for LASSO under different setups when the number of variables is of a greater order than the sample size. Candes and Tao (2007) introduce the Dantzig selector to handle the NP-dimensional variable selection problem, which was shown to behave similarly to Lasso by Bickel *et al.* (2009). Zhang (2009) is among the first to study the non-convex penalized least-squares estimator with NP-dimensionality and demonstrates its advantages over LASSO. He also develops the PLUS algorithm to find the solution path that has the desired sampling properties. Fan and Lv (2008) and Huang *et al.* (2008) introduce the independence screening procedure to reduce the dimensionality in the context of least-squares. The former establishes the sure screening property with NP-dimensionality and the latter also studies the bridge regression, a folded-concave penalty approach. Fan and Fan (2008) investigate the impact of dimensionality on ultra-high dimensional classification and establish an oracle property for features annealed independence rules. Lv and Fan (2009) make important connections between model selection and sparse recovery using folded-concave penalties and establish a nonasymptotic weak oracle property for the penalized least squares estimator with NP-dimensionality. There are also a number of important papers on establishing the oracle inequalities for penalized empirical risk minimization. For example, Bunea *et al.* (2007) establish sparsity oracle inequalities for the Lasso under quadratic loss in the context of least-squares; van de Geer (2008) obtains a nonasymptotic oracle inequality for the empirical risk minimizer with the L_1-penalty in the context of GLMs; Koltchinskii (2008) proves oracle inequalities for penalized least squares with entropy penalization.

The penalization methods are also widely used in covariance matrix estimation. This has been studied by a number of authors on the estimation of sparse covariance matrix, sparse precision matrix, and sparse Cholesky decomposition, using the Gaussian likelihood or pseudo-likelihood. See, for example, Huang *et al.* (2006), Meinshausen and Bühlmann

(2006), Levina *et al.* (2008), Rothman *et al.* (2008), and Lam and Fan (2009), among others. For these more specific models, stronger results can be obtained.

The rest of the paper is organized as follows. In Section 2, we discuss the choice of penalty functions and characterize the non-concave penalized likelihood estimator and its global optimality. We study the nonasymptotic weak oracle properties and oracle properties of non-concave penalized likelihood estimator in Sections 3 and 4, respectively. Section 5 discusses algorithms for solving regularization problems with concave penalties including the SCAD. In Section 6, we present three numerical examples using both simulated and real data sets. We provide some discussions of our results and their implications in Section 7. Proofs are presented in Section 8. Technical details are relegated to the Appendix.

2 Non-concave penalized likelihood estimation

In this section we discuss the choice of penalty functions in regularization methods and characterize the non-concave penalized likelihood estimator as well as its global optimality.

2.1 Penalty function

For any penalty function $p_\lambda(\cdot)$, we let $\rho(t;\lambda) = \lambda^{-1}p_\lambda(t)$. For simplicity, we will drop its dependence on λ and write $\rho(t;\lambda)$ as $\rho(t)$ when there is no confusion. Many penalty functions have been proposed in the literature for regularization. For example, the best subset selection amounts to using the L_0 penalty. The ridge regression uses the L_2 penalty. The L_q penalty $\rho(t) = t^q$ for $q \in (0, 2)$ bridges these two cases (Frank and Friedman, 1993). Breiman (1995) introduces the non-negative garrote for shrinkage estimation and variable selection. Lasso (Tibshirani, 1996) uses the L_1-penalized least squares. The SCAD penalty (Fan, 1997; Fan and Li, 2001) is the function whose derivative is given by

$$p'_\lambda(t) = \lambda \left\{ I\left(t \le \lambda\right) + \frac{\left(a\lambda - t\right)_+}{\left(a-1\right)\lambda} I\left(t > \lambda\right) \right\}, \quad t \ge 0, \text{ for some } a > 2, \tag{4}$$

where often $a = 3.7$ is used, and MCP (Zhang, 2009) is defined through $p'_\lambda(t) = \left(a\lambda - t\right)_+ / a$. Clearly the SCAD penalty takes off at the origin as the L_1 penalty and then levels off, and MCP translates the flat part of the derivative of SCAD to the origin. A family of folded concave penalties that bridge the L_0 and L_1 penalties were studied by Lv and Fan (2009).

Hereafter we consider penalty functions $p_\lambda(\cdot)$ that satisfy the following condition:

Condition 1. *$\rho(t;\lambda)$ is increasing and concave in $t \in [0, \infty)$, and has a continuous derivative $\rho'(t;\lambda)$ with $\rho'(0+;\lambda) > 0$. In addition, $\rho'(t;\lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+;\lambda)$ is independent of λ.*

The above class of penalty functions has been considered by Lv and Fan (2009). Clearly the L_1 penalty is a convex function that falls at the boundary of the class of penalty functions

satisfying Condition 1. Fan and Li (2001) advocate penalty functions that give estimators with three desired properties: unbiasedness, sparsity and continuity, and provide insights into them (see also Antoniadis and Fan, 2001). Both SCAD and MCP with $a \geq 1$ satisfy Condition 1 and the above three properties simultaneously. The L_1 penalty also satisfies Condition 1 as well as the sparsity and continuity, but it does not enjoy the unbiasedness, since its derivative is identically one on $[0, \infty)$ with the derivative at zero understood as the right derivative. However, our results are applicable to the L_1-penalized regression. Condition 1 is needed for establishing the oracle properties of non-concave penalized likelihood estimator.

2.2 Non-concave penalized likelihood estimator

It is generally difficult to study the global maximizer of the penalized likelihood analytically without concavity. As is common in the literature, we study the behavior of local maximizers.

We introduce some notation to simplify our presentation. For any $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T \in \mathbf{R}^n$, define

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \cdots, b'(\theta_n))^T \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}\{b''(\theta_1), \cdots, b''(\theta_n)\}. \tag{5}$$

It is known that the n-dimensional response vector $\mathbf{y}$ following the distribution in (1) has mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\phi\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$. Let $\bar{\rho}(t) = \text{sgn}(t)\rho'(|t|)$, $t \in \mathbf{R}$ and $\bar{\rho}(\mathbf{v}) = (\bar{\rho}(v_1), \cdots, \bar{\rho}(v_q))^T$, $\mathbf{v} = (v_1, \cdots, v_q)^T$, where sgn denotes the sign function. We denote by $\|\cdot\|_q$ the L_q norm of a vector or matrix for $q \in [0, \infty]$. Following Zhang (2009), define the local concavity of the penalty ρ at $\mathbf{v} = (v_1, \cdots, v_q)^T \in \mathbf{R}^q$ with $\|\mathbf{v}\|_0 = q$ as

$$\kappa(\rho; \mathbf{v}) = \lim_{\epsilon \to 0+} \max_{1 \leq j \leq q} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}. \tag{6}$$

By the concavity of ρ in Condition 1, we have $\kappa(\rho; \mathbf{v}) \geq 0$. It is easy to show by the mean-value theorem that $\kappa(\rho; \mathbf{v}) = \max_{1 \leq j \leq q} -\rho''(|v_j|)$ provided that the second derivative of ρ is continuous. For the SCAD penalty, $\kappa(\rho; \mathbf{v}) = 0$ unless some component of $|\mathbf{v}|$ takes values in $[\lambda, a\lambda]$. In the latter case, $\kappa(\rho; \mathbf{v}) = (a-1)^{-1}\lambda^{-1}$.

Throughout the paper, we use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to represent the smallest and largest eigenvalues of a symmetric matrix, respectively.

The following theorem gives a sufficient condition on the strict local maximizer of the penalized likelihood $Q_n(\boldsymbol{\beta})$ in (3) (see Lv and Fan (2009) for the case of penalized least squares).

Theorem 1 (Characterization of PMLE)**.** *Assume that p_λ satisfies Condition 1. Then $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ is a strict local maximizer of the non-concave penalized likelihood $Q_n(\boldsymbol{\beta})$ defined by*

(3) if

$$\mathbf{X}_1^T\mathbf{y} - \mathbf{X}_1^T\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) - n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\beta}}_1) = \mathbf{0}, \tag{7}$$

$$\|\mathbf{z}\|_\infty < \rho'(0+), \tag{8}$$

$$\lambda_{\min}\left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{X}_1\right] > n\lambda_n\kappa(\rho;\widehat{\boldsymbol{\beta}}_1), \tag{9}$$

where $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *respectively denote the submatrices of* $\mathbf{X}$ *formed by columns in* $\mathrm{supp}(\widehat{\boldsymbol{\beta}})$ *and its complement,* $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_1$ *is a subvector of* $\widehat{\boldsymbol{\beta}}$ *formed by all nonzero components, and* $\mathbf{z} = (n\lambda_n)^{-1}\mathbf{X}_2^T[\mathbf{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})]$. *On the other hand, if* $\widehat{\boldsymbol{\beta}}$ *is a local maximizer of* $Q_n(\boldsymbol{\beta})$, *then it must satisfy (7) – (9) with strict inequalities replaced by nonstrict inequalities.*

There is only a tiny gap (nonstrict versus strict inequalities) between the necessary condition for local maximizer and sufficient condition for strict local maximizer. Conditions (7) and (9) ensure that $\widehat{\boldsymbol{\beta}}$ is a strict local maximizer of (3) when constrained on the $\|\widehat{\boldsymbol{\beta}}\|_0$-dimensional subspace $\{\boldsymbol{\beta} \in \mathbf{R}^p : \boldsymbol{\beta}_c = \mathbf{0}\}$ of $\mathbf{R}^p$, where $\boldsymbol{\beta}_c$ denotes the subvector of $\boldsymbol{\beta}$ formed by components in the complement of $\mathrm{supp}(\widehat{\boldsymbol{\beta}})$. Condition (8) makes sure that the sparse vector $\widehat{\boldsymbol{\beta}}$ is indeed a strict local maximizer of (3) on the whole space $\mathbf{R}^p$.

When ρ is the L_1 penalty, the penalized likelihood function $Q_n(\boldsymbol{\beta})$ in (3) is concave in $\boldsymbol{\beta}$. Then the classical convex optimization theory applies to show that $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ is a global maximizer if and only if there exists a subgradient $\mathbf{z} \in \partial L_1(\widehat{\boldsymbol{\beta}})$ such that

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) - n\lambda_n\mathbf{z} = \mathbf{0}, \tag{10}$$

that is, it satisfies the Karush-Kuhn-Tucker (KKT) conditions, where the subdifferential of the L_1 penalty is given by $\partial L_1(\widehat{\boldsymbol{\beta}}) = \{\mathbf{z} = (z_1, \cdots, z_p)^T \in \mathbf{R}^p : z_j = \mathrm{sgn}(\widehat{\beta}_j)$ for $\widehat{\beta}_j \neq 0$ and $z_j \in [-1, 1]$ otherwise$\}$. Thus condition (10) reduces to (7) and (8) with strict inequality replaced by nonstrict inequality. Since $\kappa(\rho; \mathbf{v}) = 0$ for the L_1-penalty, condition (9) holds provided that $\mathbf{X}_1^T\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\mathbf{X}_1$ is nonsingular. However, to ensure that $\widehat{\boldsymbol{\beta}}$ is the strict maximizer we need the strict inequality in (8).

2.3 Global optimality

It is a natural question of when the non-concave penalized maximum likelihood estimator (NCPMLE) $\widehat{\boldsymbol{\beta}}$ is a global maximizer of the penalized likelihood $Q_n(\boldsymbol{\beta})$. We characterize such a property from two perspectives.

2.3.1 Global optimality

Assume that the $n \times p$ design matrix $\mathbf{X}$ has full column rank p. This implies that $p \leq n$. Since $b''(\theta)$ is always positive, it is easy to show that the Hessian matrix of $-\ell_n(\boldsymbol{\beta})$ is always positive definite, which entails that the log-likelihood function $\ell_n(\boldsymbol{\beta})$ is strictly concave in

$\boldsymbol{\beta}$. Thus there exists a unique maximizer $\boldsymbol{\beta}_*$ of $\ell_n(\boldsymbol{\beta})$. Let $\mathcal{L}_c = \{\boldsymbol{\beta} \in \mathbf{R}^p : \ell_n(\boldsymbol{\beta}) \geq c\}$ be a sublevel set of $-\ell_n(\boldsymbol{\beta})$ for some $c < \ell_n(\mathbf{0})$ and

$$\kappa(p_\lambda) = \sup_{t_1 < t_2 \in (0,\infty)} -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1}$$

be the maximum concavity of the penalty function p_λ. For the L_1 penalty, SCAD, and MCP, we have $\kappa(p_\lambda) = 0$, $(a-1)^{-1}$, and a^{-1}, respectively. The following proposition gives a sufficient condition on the global optimality of NCPMLE.

Proposition 1 (Global optimality)**.** *Assume that* $\mathbf{X}$ *has rank* p *and satisfies*

$$\min_{\boldsymbol{\beta} \in \mathcal{L}_c} \lambda_{\min}\left[n^{-1}\mathbf{X}^T \boldsymbol{\Sigma}\left(\mathbf{X}\boldsymbol{\beta}\right)\mathbf{X}\right] \geq \kappa(p_{\lambda_n}). \tag{11}$$

Then the NCPMLE $\widehat{\boldsymbol{\beta}}$ *is a global maximizer of the penalized likelihood* $Q_n(\boldsymbol{\beta})$ *if* $\widehat{\boldsymbol{\beta}} \in \mathcal{L}_c$*.*

Note that for penalized least-squares, (11) reduces to

$$\lambda_{\min}\left(n^{-1}\mathbf{X}^T\mathbf{X}\right) \geq \kappa(p_{\lambda_n}). \tag{12}$$

This condition holds for sufficiently large a in SCAD and MCP, when the correlation between covariates is not too strong. The latter holds for design matrices constructed by using spline bases to approximate a nonparametric function. According to Proposition 1, under (12), the penalized least-squares with folded-concave penalty is a global minimum.

The proposition below gives a condition under which the penalty term in (3) does not change the global maximizer. It will be used to derive the condition under which the PMLE is the same as the oracle estimator in Proposition 3(b). Here for simplicity we consider the SCAD penalty p_λ given by (4), and the technical arguments are applicable to other folded-concave penalties as well.

Proposition 2 (Robustness)**.** *Assume that* $\mathbf{X}$ *has rank* p *with* $p = s$ *and there exists some* $c < \ell_n(\mathbf{0})$ *such that* $\min_{\boldsymbol{\beta} \in \mathcal{L}_c} \lambda_{\min}[n^{-1}\mathbf{X}^T\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}] \geq c_0$ *for some* $c_0 > 0$*. Then the SCAD penalized likelihood estimator* $\widehat{\boldsymbol{\beta}}$ *is the global maximizer and equals* $\boldsymbol{\beta}_*$ *if* $\widehat{\boldsymbol{\beta}} \in \mathcal{L}_c$ *and* $\min_{j=1}^p |\widehat{\beta}_j| > (a + \frac{1}{2c_0})\lambda_n$*, where* $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$*.*

2.3.2 Restricted global optimality

When $p > n$, it is hard to show the global optimality of a local maximizer. However, we can study the global optimality of the NCPMLE $\widehat{\boldsymbol{\beta}}$ on the union of coordinate subspaces. A subspace of $\mathbf{R}^p$ is called coordinate subspace if it is spanned by a subset of the natural basis $\{\mathbf{e}_1, \cdots, \mathbf{e}_p\}$, where each $\mathbf{e}_j$ is the p-vector with j-th component 1 and 0 elsewhere. Here each $\mathbf{e}_j$ corresponds to the j-th predictor $\mathbf{x}_j$. We will investigate the global optimality of $\widehat{\boldsymbol{\beta}}$ on the union $\mathbb{S}_s$ of all s-dimensional coordinate subspaces of $\mathbf{R}^p$ in Proposition 3(a).

Of particularly interest is to derive the conditions under which the PMLE is also an oracle estimator, in addition to possessing the above restricted global optimal estimator on $\mathbb{S}_s$. To this end, we introduce an identifiability condition on the true model $\text{supp}(\boldsymbol{\beta}_0)$. The true model is called δ-identifiable for some $\delta > 0$ if

$$\max_{\boldsymbol{\beta} \in \mathcal{A}_0} \ell_n(\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta} \in \mathbb{S}_s \setminus \mathcal{A}_0} \ell_n(\boldsymbol{\beta}) \geq \delta, \tag{13}$$

where $\mathcal{A}_0 = \{(\beta_1, \cdots, \beta_p)^T \in \mathbf{R}^p : \beta_j = 0 \text{ for } j \notin \text{supp}(\boldsymbol{\beta}_0)\}$. In other words, $\text{supp}(\boldsymbol{\beta}_0)$ is the best subset of size s, with a margin at least δ. The following proposition is an easy consequence of Propositions 1 and 2.

Proposition 3 (Global optimality on $\mathbb{S}_s$)**.**

a) If the conditions of Proposition 1 are satisfied for each $n \times (2s)$ submatrix of $\mathbf{X}$, then the NCPMLE $\widehat{\boldsymbol{\beta}}$ is a global maximizer of $Q_n(\boldsymbol{\beta})$ on $\mathbb{S}_s$.

b) Assume that the conditions of Proposition 2 are satisfied for the $n \times s$ submatrix of $\mathbf{X}$ formed by columns in $\text{supp}(\boldsymbol{\beta}_0)$, the true model is δ-identifiable for some $\delta > \frac{(a+1)s\lambda_n^2}{2}$, and $\text{supp}(\widehat{\boldsymbol{\beta}}) = \text{supp}(\boldsymbol{\beta}_0)$. Then the SCAD penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$ is the global maximizer on $\mathbb{S}_s$ and equals to the oracle maximum likelihood estimator $\boldsymbol{\beta}_*$.

On the event that the PMLE estimator is the same as the oracle estimator, it possesses of course the oracle property.

3 Nonasymptotic weak oracle properties

In this section we study a nonasymptotic property of the non-concave penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$, called the weak oracle property introduced by Lv and Fan (2009) in the setting of penalized least squares. The weak oracle property means sparsity in the sense of $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$, and consistency under the L_∞ loss, where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ and $\widehat{\boldsymbol{\beta}}_1$ is a subvector of $\widehat{\boldsymbol{\beta}}$ formed by components in $\text{supp}(\boldsymbol{\beta}_0) = \{1, \cdots, s\}$. This property is weaker than the oracle property introduced by Fan and Li (2001).

3.1 Regularity conditions

As mentioned before, we condition on the design matrix $\mathbf{X}$ and use the p_λ penalty in the class satisfying Condition 1. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively be the submatrices of the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ formed by columns in $\text{supp}(\boldsymbol{\beta}_0)$ and its complement, and $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0$. To simplify the presentation, we assume without loss of generality that each covariate $\mathbf{x}_j$ has been standardized so that $\|\mathbf{x}_j\|_2 = \sqrt{n}$. If the covariates have not been standardized, the results still hold with $\|\mathbf{x}_j\|_2$ assumed to be in the order of $\sqrt{n}$. Let

$$d_n = 2^{-1} \min \{|\beta_{0,j}| : \beta_{0,j} \neq 0\} \tag{14}$$

be half of the minimum signal. We make the following assumptions on the design matrix and the distribution of the response.

Let $\{b_s\}$ be a diverging sequence of positive numbers that depends on the nonsparsity size s and hence depends on n. Recall that $\boldsymbol{\beta}_1$ is the non-vanishing components of the true parameter $\boldsymbol{\beta}_0$.

Condition 2. *The design matrix* $\mathbf{X}$ *satisfies*

$$\left\| \left[\mathbf{X}_1^T \boldsymbol{\Sigma} \left(\boldsymbol{\theta}_0 \right) \mathbf{X}_1 \right]^{-1} \right\|_\infty = O(b_s n^{-1}), \tag{15}$$

$$\left\| \mathbf{X}_2^T \boldsymbol{\Sigma} \left(\boldsymbol{\theta}_0 \right) \mathbf{X}_1 \left[\mathbf{X}_1^T \boldsymbol{\Sigma} \left(\boldsymbol{\theta}_0 \right) \mathbf{X}_1 \right]^{-1} \right\|_\infty \leq \min \left\{ C \frac{\rho'(0+)}{\rho'(d_n)}, O(n^{\alpha_1}) \right\}, \tag{16}$$

$$\max_{\boldsymbol{\delta} \in \mathcal{N}_0} \max_{j=1}^p \lambda_{\max} \left[\mathbf{X}_1^T \operatorname{diag} \left\{ |\mathbf{x}_j| \circ \left| \boldsymbol{\mu}'' \left(\mathbf{X}_1 \boldsymbol{\delta} \right) \right| \right\} \mathbf{X}_1 \right] = O(n), \tag{17}$$

where the L_∞ *norm of a matrix is the maximum of the* L_1 *norm of each row,* $C \in (0,1)$*,* $\alpha_1 \in [0, 1/2]$*,* $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_\infty \leq d_n\}$*, the derivative is taken componentwise, and* $\circ$ *denotes the Hadamard (componentwise) product.*

Here and below, ρ is associated with regularization parameter λ_n satisfying (18) unless specified otherwise. For the classical Gaussian linear regression model, we have $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = I_n$. In this case, since we will assume that $s \ll n$, condition (15) usually holds with $b_s = 1$ if the covariates are nearly uncorrelated. In fact, Wainwright (2009) shows that $\|[\mathbf{X}_1^T \mathbf{X}_1]^{-1}\|_\infty = O_P(n^{-1})$ if the rows of $\mathbf{X}_1$ are i.i.d. Gaussian vectors with $\|[E\mathbf{X}_1^T \mathbf{X}_1]^{-1}\|_\infty = O_P(n^{-1})$. In general, since

$$\|[\mathbf{X}_1^T \mathbf{X}_1]^{-1}\|_\infty \leq \sqrt{s} / \lambda_{\min}(\mathbf{X}_1^T \mathbf{X}_1),$$

we can take $b_s = s^{1/2}$ if $\lambda_{\min}(\mathbf{X}_1^T \mathbf{X}_1)^{-1} = O(n^{-1})$. More generally, (15) can be bounded as

$$\left\| \left[\mathbf{X}_1^T \boldsymbol{\Sigma} \left(\boldsymbol{\theta}_0 \right) \mathbf{X}_1 \right]^{-1} \right\|_\infty = d^{-1} \left\| \left[\mathbf{X}_{1,S}^T \mathbf{X}_{1,S} \right]^{-1} \right\|_\infty$$

and the above remark for the multiple regression model applies to the submatrix $\mathbf{X}_{1,S}$, which consists of rows of the samples with $b''(\theta_i) > d$ for some $d > 0$.

The left hand side of (16) is the multiple regression coefficients of each unimportant variable in $\mathbf{X}_2$ on $\mathbf{X}_1$, using the weighted least squares with weights $\{b''(\theta_i)\}$. Condition (16) controls the uniform growth rate of the L_1-norm of these multiple regression coefficients, a notion of weak correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$. If each element of the multiple regression coefficients is of order $O(1)$, then the L_1 norm is of order $O(s)$. Hence, we can handle the non-sparse dimensionality $s = O(n^{\alpha_1})$, by (16), as long as the first term in (16) dominates, which occurs for SCAD type of penalty with $d_n \gg \lambda_n$. Of course, the actual dimensionality can be higher or lower, depending on the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$, but for finite non-sparse dimensionality $s = O(1)$, (16) is usually satisfied. When a folded-concave penalty is used,

the upper bound on the right hand side of (16) can grow to ∞ at rate $O(n^{\alpha_1})$. In contrast, when the L_1 penalty is used, the upper bound in (16) is more restrictive, requiring uniformly less than 1. This condition is the same as the strong irrepresentable condition of Zhao and Yu (2006) for the consistency of the LASSO estimator, namely $\|\mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\|_\infty \le C$. It is a drawback of the L_1 penalty.

For the Gaussian linear regression model, condition (17) holds automatically.

We now choose the regularization parameter λ_n and introduce Condition 3. We will assume that half of the minimum signal $d_n \ge n^{-\gamma}\log n$ for some $\gamma \in (0, 1/2]$. Take λ_n satisfying

$$p'_{\lambda_n}(d_n) = o(b_s^{-1}n^{-\gamma}\log n) \quad \text{and} \quad \lambda_n \gg n^{-\alpha}(\log n)^2, \tag{18}$$

where $\alpha = \min(\frac{1}{2}, 2\gamma - \alpha_0) - \alpha_1$ and b_s is associated with the nonsparsity size $s = O(n^{\alpha_0})$.

Condition 3. *Assume that $d_n \ge n^{-\gamma}\log n$ and $b_s = o\{\min(n^{1/2-\gamma}\sqrt{\log n}, s^{-1}n^{\gamma}/\log n)\}$. In addition, assume that λ_n satisfies (18) and $\lambda_n\kappa_0 = o(\tau_0)$, where $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho;\boldsymbol{\delta})$ and $\tau_0 = \min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}[n^{-1}\mathbf{X}_1^T\boldsymbol{\Sigma}(\mathbf{X}_1\boldsymbol{\delta})\mathbf{X}_1]$, and that $\max_{j=1}^p\|\mathbf{x}_j\|_\infty = o(n^{\alpha}/\sqrt{\log n})$ if the responses are unbounded.*

The condition that $\lambda_n\kappa_0 = o(\tau_0)$, is needed to ensure condition (9). The condition always holds when $\kappa_0 = 0$ and is satisfied for the SCAD type of penalty when $d_n \gg \lambda_n$.

In view of (7) and (8), to study the non-concave penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$ we need to analyze the deviation of the p-dimensional random vector $\mathbf{X}^T\mathbf{Y}$ from its mean $\mathbf{X}^T\boldsymbol{\mu}(\boldsymbol{\theta}_0)$, where $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ denotes the n-dimensional random response vector in the GLM (1). The following proposition, whose proof is given in Section 8.5, characterizes such deviation for the case of bounded responses and the case of unbounded responses satisfying a moment condition, respectively.

Proposition 4 (Deviation)**.** *Let $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ be the n-dimensional independent random response vector and $\mathbf{a} \in \mathbf{R}^n$. Then*

a) If $Y_1, \cdots, Y_n$ are bounded in $[c, d]$ for some $c, d \in \mathbf{R}$, then for any $\varepsilon \in (0, \infty)$,

$$P\left(\left|\mathbf{a}^T\mathbf{Y} - \mathbf{a}^T\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right| > \varepsilon\right) \le 2\exp\left[-\frac{2\varepsilon^2}{\|\mathbf{a}\|_2^2\,(d-c)^2}\right]. \tag{19}$$

b) If $Y_1, \cdots, Y_n$ are unbounded and there exist some $M, v_0 \in (0, \infty)$ such that

$$\max_{i=1,\cdots,n} E\left\{\exp\left[\frac{|Y_i - b'\left(\theta_{0,i}\right)|}{M}\right] - 1 - \frac{|Y_i - b'\left(\theta_{0,i}\right)|}{M}\right\}M^2 \le \frac{v_0}{2} \tag{20}$$

with $(\theta_{0,1}, \cdots, \theta_{0,n})^T = \boldsymbol{\theta}_0$, then for any $\varepsilon \in (0, \infty)$,

$$P\left(\left|\mathbf{a}^T\mathbf{Y} - \mathbf{a}^T\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right| > \varepsilon\right) \le 2\exp\left[-\frac{1}{2}\frac{\varepsilon^2}{\|\mathbf{a}\|_2^2\,v_0 + \|\mathbf{a}\|_\infty M\varepsilon}\right]. \tag{21}$$

In light of (1), it is known that for the exponential family, the moment-generating function of Y_i is given by

$$E\exp\left\{t\left[Y_i - b'\left(\theta_{0,i}\right)\right]\right\} = \exp\left\{\phi^{-1}\left[b\left(\theta_{0,i}+t\phi\right) - b\left(\theta_{0,i}\right) - b'\left(\theta_{0,i}\right)t\phi\right]\right\},$$

where $\theta_{0,i} + t\phi$ is in the domain of $b(\cdot)$. Thus the moment condition (20) is reasonable. It is easy to show that condition (20) holds for the Gaussian linear regression model and for the Poisson regression model with bounded mean responses. Similar probability bounds also hold for sub-Gaussian errors.

We now express the results in Proposition 4 in a unified form. For the case of bounded responses, we define $\varphi(\varepsilon) = 2e^{-c_1\varepsilon^2}$ for $\varepsilon \in (0,\infty)$, where $c_1 = 2/(d-c)^2$. For the case of unbounded responses satisfying the moment condition (20), we define $\varphi(\varepsilon) = 2e^{-c_1\varepsilon^2}$, where $c_1 = 1/(2v_0 + 2M)$. Then the exponential bounds in (19) and (21) can be expressed as

$$P\left(\left|\mathbf{a}^T\mathbf{Y} - \mathbf{a}^T\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right| > \|\mathbf{a}\|_2\,\varepsilon\right) \le \varphi(\varepsilon), \tag{22}$$

where $\varepsilon \in (0,\infty)$ if the responses are bounded and $\varepsilon \in (0, \|\mathbf{a}\|_2/\|\mathbf{a}\|_\infty]$ if the responses are unbounded.

3.2 Weak oracle properties

Theorem 2 (Weak oracle property)**.** *Assume that Conditions 1–3 and the probability bound (22) are satisfied, $s = o(n)$, and $\log p = O(n^{1-2\alpha})$. Then there exists a non-concave penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$ such that for sufficiently large n, with probability at least $1 - 2[sn^{-1} + (p-s)e^{-n^{1-2\alpha}\log n}]$, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ satisfies:*

a) (Sparsity). $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$*;*

b) (L_∞ loss). $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_\infty = O(n^{-\gamma}\log n)$*,*

where $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_1$ are respectively the subvectors of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ formed by components in $\operatorname{supp}(\boldsymbol{\beta}_0)$.

Under the given regularity conditions, the dimensionality p is allowed to grow up to exponentially fast with the sample size n. The growth rate of $\log p$ is controlled by $1 - 2\alpha$. It also enters the nonasymptotic probability bound. This probability tends to 1 under our technical assumptions. From the proof of Theorem 2, we see that with asymptotic probability one, the L_∞ estimation loss of the non-concave penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$ is bounded from above by three terms (see (45)), where the second term $b_s\lambda_n\rho'(d_n)/\rho'(0+)$ is associated with the penalty function ρ. For the L_1 penalty, the ratio $\rho'(d_n)/\rho'(0+)$ is equal to one, and for other concave penalties, it can be (much) smaller than one. This is in line with the fact shown by Fan and Li (2001) that concave penalties can reduce the biases of estimates.

Under the specific setting of penalized least squares, the above weak oracle property is slightly different from that of Lv and Fan (2009).

The value of γ can be taken as large as $1/2$ for concave penalties. In this case, the dimensionality that the penalized least-squares can handle is as high as $\log p = O(n^{2\alpha_1})$ when $\alpha_0 \leq 1/2$, which is usually smaller than that for the case of $\gamma < \frac{1}{4} + \frac{\alpha_0}{2}$. The large value of γ puts more stringent condition on the design matrix. To see this, Condition 3 entails that $b_s = o(\sqrt{\log n})$ and hence (15) becomes tighter.

In the classical setting of $\gamma = 1/2$, the consistency rate of $\widehat{\boldsymbol{\beta}}$ under the L_2 norm becomes $O_P(\sqrt{s}n^{-1/2}\log n)$, which is slightly slower than $O_P(\sqrt{s}n^{-1/2})$. This is because it is derived by using the L_∞ loss of $\widehat{\boldsymbol{\beta}}$ in Theorem 2b). The use of the L_∞ norm is due to the technical difficulty of proving the existence of a solution to the nonlinear equation (7).

3.3 Sampling properties of L_1-based PMLE

When the L_1-penalty is applied, the penalized likelihood $Q_n(\boldsymbol{\beta})$ in (3) is concave. The local maximizer in Theorems 1 and 2 becomes the global maximizer. Due to its popularity, we now examine the implications of Theorem 2 in the context of penalized least-squares and penalized likelihood.

For the penalized least-squares, Condition 2 becomes

$$\left\| \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \right\|_\infty = O(b_s n^{-1}), \tag{23}$$

$$\left\| \mathbf{X}_2^T \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{X}_1 \right)^{-1} \right\|_\infty \leq C < 1. \tag{24}$$

Condition (17) holds automatically and Condition (18) becomes

$$\lambda_n = o(b_s^{-1} n^{-\gamma} \log n) \quad \text{and} \quad \lambda_n \gg n^{-\alpha} (\log n)^2. \tag{25}$$

As a corollary of Theorem 2, we have

Corollary 1 (Penalized L_1 estimator)**.** *Under Conditions 2 and 3 and probability bound (22), if $s = o(n)$ and $\log p = O(n^{1-2\alpha})$, then the penalized L_1 likelihood estimator $\widehat{\boldsymbol{\beta}}$ has model selection consistency with rate $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_\infty = O(n^{-\gamma}\log n)$.*

For the penalized least-squares, Corollary 1 continues to hold without normality assumption, as long as probability bound (22) holds. In this case, the result is stronger than that of Zhao and Yu (2006) and Lv and Fan (2009).

4 Oracle properties

In this section we study the oracle property (Fan and Li, 2001) of the non-concave penalized likelihood estimator $\widehat{\boldsymbol{\beta}}$. We assume that the nonsparsity size $s \ll n$ and the dimensionality

satisfies $\log p = O(n^{\alpha})$ for some $\alpha \in (0, 1/2)$, which is related to the notation in Section 3. We impose the following regularity conditions.

Condition 4. *The design matrix* $\mathbf{X}$ *satisfies*

$$\min_{\boldsymbol{\delta}\in\mathcal{N}_0} \lambda_{\min}\left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\mathbf{X}_1\boldsymbol{\delta}\right)\mathbf{X}_1\right] \geq cn, \quad \mathrm{tr}[\mathbf{X}_1^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}_1] = O(sn), \tag{26}$$

$$\left\|\mathbf{X}_2^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\right\|_{2,\infty} = O(n), \tag{27}$$

$$\max_{\boldsymbol{\delta}\in\mathcal{N}_0} \max_{j=1}^{p} \lambda_{\max}\left[\mathbf{X}_1^T \mathrm{diag}\left\{|\mathbf{x}_j| \circ \left|\boldsymbol{\mu}''\left(\mathbf{X}_1\boldsymbol{\delta}\right)\right|\right\}\mathbf{X}_1\right] = O(n), \tag{28}$$

where $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_\infty \leq d_n\}$, *c is some positive constant, and* $\|\mathbf{B}\|_{2,\infty} = \max_{\|\mathbf{v}\|_2=1} \|\mathbf{B}\mathbf{v}\|_\infty$.

Condition 5. *Assume that* $d_n \gg \lambda_n \gg \max\{(s/n)^{1/2}, n^{(\alpha-1)/2}(\log n)^{1/2}\}$, $p'_{\lambda_n}(d_n) = O(n^{-1/2})$, *and* $\lambda_n\kappa_0 = o(1)$, *where* $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho;\boldsymbol{\delta})$, *and in addition that* $\max_{j=1}^{p}\|\mathbf{x}_j\|_\infty = o(n^{\frac{1-\alpha}{2}}/\sqrt{\log n})$ *if the responses are unbounded.*

Condition 4 is generally stronger than Condition 2. In fact, by $d_n \gg \lambda_n$ in Condition 5, the first condition in (16) holds automatically for SCAD type of penalties, since $p'_{\lambda_n}(d_n) = 0$ when n is large enough. Thus Condition 5 is less restrictive for SCAD-like penalties, since $\kappa_0 = 0$ for sufficiently large n.

However, for the L_1 penalty, $\lambda_n = p'_{\lambda_n}(d_n) = O(n^{-1/2})$ is incompatible with $\lambda_n \gg (s/n)^{1/2}$. This suggests that the L_1 penalized likelihood estimator generally cannot achieve the consistency rate of $O_P(\sqrt{s}n^{-1/2})$ established in Theorem 3 and does not have the oracle property established in Theorem 4, when the dimensionality p is diverging with the sample size n. In fact, this problem was observed by Fan and Li (2001) and proved by Zou (2006) even for finite p. It still persists with growing dimensionality.

We now state the existence of the NCPMLE and its rate of convergence. It improves the rate results given by Theorem 2.

Theorem 3 (Existence of non-concave penalized likelihood estimator)**.** *Assume that Conditions 1, 4 and 5 and the probability bound (22) hold. Then there exists a strict local maximizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ *of the penalized likelihood* $Q_n(\boldsymbol{\beta})$ *such that* $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ *with probability tending to 1 as* $n \to \infty$ *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_P(\sqrt{s}n^{-1/2})$, *where* $\widehat{\boldsymbol{\beta}}_1$ *is a subvector of* $\widehat{\boldsymbol{\beta}}$ *formed by components in* $\mathrm{supp}(\boldsymbol{\beta}_0)$.

Theorem 3 can be thought of as answering the question that given the dimensionality, how strong the minimum signal d_n should be in order for the penalized likelihood estimator to have some nice properties, through Conditions 4 and 5. On the other hand, Theorem 2 can be thought of as answering the question that given the strength of the minimum signal d_n, how high dimensionality the penalized likelihood methods can handle, through Conditions 2 and 3. While the details are different, these conditions are related.

To establish the asymptotic normality, we need additional condition, which is related to the Lyapunov condition.

Condition 6. *Assume that* $p'_{\lambda_n}(d_n) = o(s^{-1/2}n^{-1/2})$, $\max_{i=1}^n E|Y_i - b'(\theta_{0,i})|^3 = O(1)$, *and* $\sum_{i=1}^n (\mathbf{z}_i^T \mathbf{B}_n^{-1} \mathbf{z}_i)^{3/2} \to 0$ *as* $n \to \infty$, *where* $(Y_1, \cdots, Y_n)^T$ *denotes the* n*-dimensional random response vector,* $(\theta_{0,1}, \cdots, \theta_{0,n})^T = \boldsymbol{\theta}_0$, $\mathbf{B}_n = \mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1$, *and* $\mathbf{X}_1 = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T$.

Theorem 4 (Oracle property)**.** *Under the conditions of Theorem 3, if Condition 6 holds and* $s = o(n^{1/3})$, *then with probability tending to 1 as* $n \to \infty$, *the non-concave penalized likelihood estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ *in Theorem 3 must satisfy:*

a) (Sparsity). $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

b) (Asymptotic normality).

$$\mathbf{A}_n \left[\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1\right]^{1/2} \left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \phi \mathbf{G}),$$

where $\mathbf{A}_n$ *is a* $q \times s$ *matrix such that* $\mathbf{A}_n \mathbf{A}_n^T \to \mathbf{G}$, $\mathbf{G}$ *is a* $q \times q$ *symmetric positive definite matrix, and* $\widehat{\boldsymbol{\beta}}_1$ *is a subvector of* $\widehat{\boldsymbol{\beta}}$ *formed by components in* $\operatorname{supp}(\boldsymbol{\beta}_0)$.

From the proof of Theorem 4, we see that for the Gaussian linear regression model, the additional restriction of $s = o(n^{1/3})$ can be relaxed, since the term in (28) vanishes in this case.

5 Implementation

In this section, we discuss algorithms for maximizing the penalized likelihood $Q_n(\boldsymbol{\beta})$ in (3) with concave penalties including the SCAD. Efficient algorithms for maximizing non-concave penalized likelihood include the LQA proposed by Fan and Li (2001) and LLA introduced by Zou and Li (2008). The coordinate optimization algorithm was used by Fu (1998) and Daubechies *et al.* (2004) for penalized least-squares with L_q-penalty. This algorithm can also be applied to optimize the group Lasso (Antoniadis and Fan, 2001; Yuan and Lin, 2006) as shown in Meier *et al.* (2008) and the penalized precision matrix estimation in Friedman *et al.* (2007).

In this paper we employ a path-following algorithm, called the iterative coordinate ascent (ICA) algorithm. Coordinate optimization type algorithms are especially appealing for large scale problems with both n and p large. It successively maximizes $Q_n(\boldsymbol{\beta})$ for regularization parameter λ in a decreasing order. ICA uses the Gauss-Seidel method, i.e., maximizing one coordinate at a time with successive displacements. Specifically, for each coordinate within each iteration, ICA uses the second order approximation of $\ell_n(\boldsymbol{\beta})$ at the p-vector from the previous step along that coordinate and maximizes the univariate penalized quadratic

approximation. It updates each coordinate if the maximizer of the corresponding univariate penalized quadratic approximation makes $Q_n(\boldsymbol{\beta})$ strictly increase. Therefore, ICA algorithm enjoys the ascent property, i.e., the resulting sequence of Q_n values is increasing for a fixed λ.

When $\ell_n(\boldsymbol{\beta})$ is quadratic in $\boldsymbol{\beta}$, e.g., for the Gaussian linear regression model, the second order approximation in ICA is exact at each step. For any $\boldsymbol{\delta} \in \mathbf{R}^p$ and $j \in \{1, \cdots, p\}$, we denote by $\widetilde{\ell}_n(\boldsymbol{\beta}; \boldsymbol{\delta}, j)$ the second order approximation of $\ell_n(\boldsymbol{\beta})$ at $\boldsymbol{\delta}$ along the j-th component, and

$$\widetilde{Q}_n(\beta_j; \boldsymbol{\delta}, j) = \widetilde{\ell}_n(\boldsymbol{\beta}; \boldsymbol{\delta}, j) - \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{29}$$

where the subvector of $\boldsymbol{\beta}$ with components in $\{1, \cdots, p\} \setminus \{j\}$ is identical to that of $\boldsymbol{\delta}$. Clearly maximizing $\widetilde{Q}_n(\cdot; \boldsymbol{\delta}, j)$ is a univariate penalized least squares problem, which admits analytical solution for many commonly used penalty functions. See the Appendix for formulae for three popular GLMs.

Pick $\lambda_{\max} \in (0, \infty)$ sufficiently large such that the maximizer of $Q_n(\boldsymbol{\beta})$ with $\lambda = \lambda_{\max}$ is $\mathbf{0}$, a decreasing sequence of regularization parameters $\{\lambda_1, \cdots, \lambda_K\}$ with $\lambda_1 = \lambda_{\max}$, and the number of iterations L.

ICA ALGORITHM.

1. Set $k = 1$ and initialize $\widehat{\boldsymbol{\beta}}^{\lambda_0} = \mathbf{0}$.
2. Initialize $\widehat{\boldsymbol{\beta}}^{\lambda_k} = \widehat{\boldsymbol{\beta}}^{\lambda_{k-1}}$, and set $S = \{1, \cdots, p\}$ and $\ell = 1$.
3. Successively for $j \in S$, let $\widehat{\beta}_j$ be the maximizer of $\widetilde{Q}_n(\beta_j; \widehat{\boldsymbol{\beta}}^{\lambda_k}, j)$, and update the j-th component of $\widehat{\boldsymbol{\beta}}^{\lambda_k}$ as $\widehat{\beta}_j$ if the updated $\widehat{\boldsymbol{\beta}}^{\lambda_k}$ strictly increases $Q_n(\boldsymbol{\beta})$. Set $S \leftarrow \mathrm{supp}(\widehat{\boldsymbol{\beta}}^{\lambda_k}) \cup \{j : |z_j| > \rho'(0+)\}$ and $\ell \leftarrow \ell + 1$, where $(z_1, \cdots, z_p)^T = (n\lambda_k)^{-1}\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k})]$.
4. Repeat Step 3 until convergence or $\ell = L + 1$. Set $k \leftarrow k + 1$.
5. Repeat Steps 2–4 until $k = K + 1$. Return p-vectors $\widehat{\boldsymbol{\beta}}^{\lambda_1}, \cdots, \widehat{\boldsymbol{\beta}}^{\lambda_K}$.

When we decrease the regularization parameter from λ_k to λ_{k+1}, using $\widehat{\boldsymbol{\beta}}^{\lambda_k}$ as an initial value for $\widehat{\boldsymbol{\beta}}^{\lambda_{k+1}}$ can speed up the convergence. The set S is introduced in Step 3 to reduce the computational cost. It is optional to add $\{j : |z_j| > \rho'(0+)\}$ to the set S in this step. In practice, we can set a small tolerance level for convergence. We can also set a level of sparsity for early stopping if desired models are only those with size up to a certain level. When the L_1 penalty is used, it is known that the choice of $\lambda = n^{-1}\|\mathbf{X}^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{0})]\|_\infty$ ensures that $\mathbf{0}$ is the global maximizer of (3). In practice, we can use this value as a proxy for $\lambda_{\max}$. We give the formulas for three commonly used GLMs and the univariate SCAD penalized least squares solution in Sections A.1 and A.2 in the Appendix, respectively.

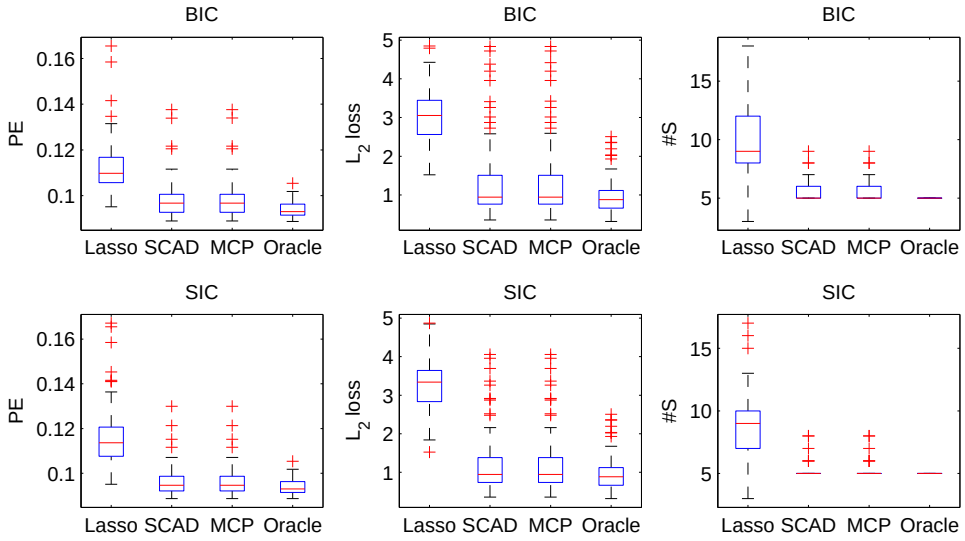

Figure 1: Boxplots of PE, L_2 loss, and #S over 100 simulations for all methods in logistic regression, where $p = 25$. The x-axis represents different methods. Top panel is for BIC and bottom panel is for SIC.

6 Numerical examples

6.1 Logistic regression

In this example, we demonstrate the performance of non-concave penalized likelihood methods in logistic regression. The data were generated from the logistic regression model (1). We set $(n, p) = (200, 25)$ and chose the true regression coefficients vector $\boldsymbol{\beta}_0$ by setting $\boldsymbol{\beta}_1 = (2.5, -1.9, 2.8, -2.2, 3)^T$. The number of simulations was 100. For each simulated data set, the rows of $\mathbf{X}$ were sampled as i.i.d. copies from $N(\mathbf{0}, \Sigma_0)$ with $\Sigma_0 = (0.5^{|i-j|})_{i,j=1,\cdots,p}$, and the response vector $\mathbf{y}$ was generated independently from the Bernoulli distribution with conditional success probability vector $g(\mathbf{X}\boldsymbol{\beta}_0)$, where $g(x) = e^x/(1+e^x)$. We compared Lasso (L_1 penalty), SCAD and MCP with the oracle estimator, all of which were implemented by the ICA algorithm to produce the solution paths. The regularization parameter λ was selected by BIC and the semi-Bayesian information criterion (SIC) introduced by Lv and Liu (2008).

Six performance measures were used to compare the methods. The first measure is the prediction error (PE) defined as $E[Y - g(\mathbf{X}^T\widehat{\boldsymbol{\beta}})]^2$, where $\widehat{\boldsymbol{\beta}}$ is the estimated coefficients vector by a method and $(\mathbf{X}^T, Y)$ is an independent test point. The second and third measures are the L_2 loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and L_1 loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$. The fourth measure is the deviance of the fitted model. The fifth measure, #S, is the number of selected variables in the final model by a method in a simulation. The sixth one, FN, measures the number of missed true variables

Table 1: Medians and robust standard deviations (in parentheses) of PE, L_2 loss, L_1 loss, deviance, #S, and FN over 100 simulations for all methods in logistic regression by BIC and SIC, where $p = 25$

Method	Measures	Lasso	SCAD	MCP	Oracle
BIC	PE	0.110(0.008)	0.097(0.006)	0.097(0.006)	0.093(0.004)
	L_2 loss	3.055(0.656)	0.943(0.550)	0.943(0.550)	0.880(0.339)
	L_1 loss	7.247(1.095)	1.867(1.461)	1.867(1.461)	1.732(0.767)
	Deviance	129.36(19.20)	111.82(15.80)	111.82(15.80)	113.12(16.05)
	#S	9(2.97)	5(0.74)	5(0.74)	5(0)
	FN	0(0)	0(0)	0(0)	0(0)
SIC	PE	0.114(0.010)	0.095(0.005)	0.095(0.005)	0.093(0.004)
	L_2 loss	3.342(0.600)	0.943(0.476)	0.943(0.476)	0.880(0.339)
	L_1 loss	7.646(1.114)	1.799(1.006)	1.799(1.006)	1.732(0.767)
	Deviance	134.93(18.35)	112.22(16.30)	112.22(16.30)	113.12(16.05)
	#S	9(2.22)	5(0)	5(0)	5(0)
	FN	0(0)	0(0)	0(0)	0(0)

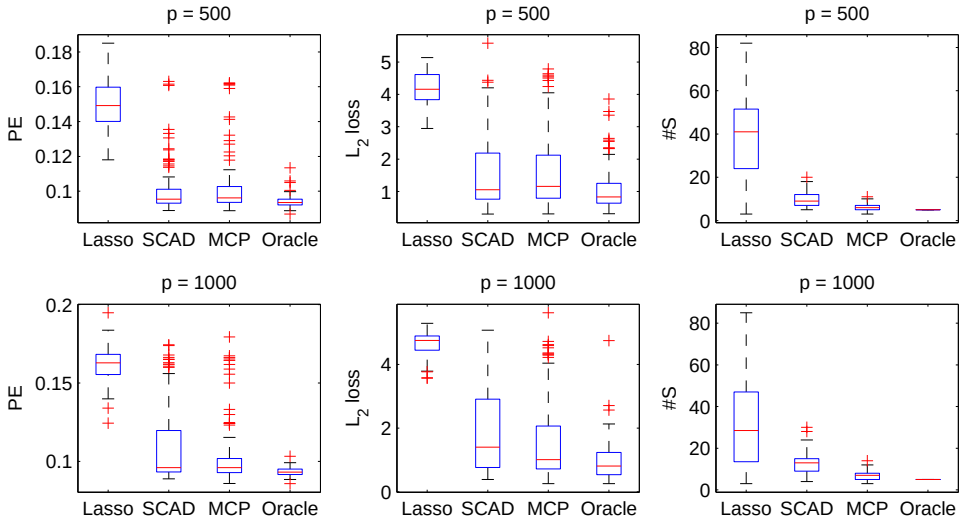

Figure 2: Boxplots of PE, L_2 loss, and #S over 100 simulations for all methods in logistic regression, where $p = 500$ and 1000. The x-axis represents different methods. Top panel is for $p = 500$ and bottom panel is for $p = 1000$.

by a method in a simulation.

In the calculation of PE, an independent test sample of size 10,000 was generated to compute the expectation. For both BIC and SIC, Lasso had median FN = 0 with some nonzeros, and SCAD and MCP had FN = 0 over 100 simulations. Table 1 and Figure 1 summarize the comparison results given by PE, L_2 loss, L_1 loss, deviance, #S, and FN, respectively for BIC and SIC. The Lasso selects larger model sizes than SCAD and MCP. Its associated median losses are also larger.

We also examined the performance of non-concave penalized likelihood methods in high dimensional logistic regression. The setting of this simulation is the same as above, except that $p = 500$ and 1000. Since p is larger than n, the information criteria break down in the tuning of λ due to the overfitting. Thus we used five-fold cross-validation based on prediction error to select the tuning parameter. Lasso had many nonzeros of FN, and SCAD and MCP had FN = 0 over almost all 100 simulations except very few nonzeros. Table 2 and Figure 2 report the comparison results given by PE, L_2 loss, L_1 loss, deviance, #S, and FN.

It is clear from Table 2 that LASSO selects far larger model size than SCAD and MCP. This is due to the bias of the L_1 penalty. The larger bias in LASSO forces the cross-validation to choose a smaller value of λ to reduce its contribution to PE. But, a smaller value of λ allows more false positive variables to be selected. The problem is certainly less severe for the SCAD penalty and MCP. The performance between SCAD and MCP is comparable, as expected.

6.2 Poisson regression

In this example, we demonstrate the performance of non-concave penalized likelihood methods in Poisson regression. The data were generated from the Poisson regression model (1). The setting of this example is similar to that in Section 6.1. We set $(n, p) = (200, 25)$ and chose the true regression coefficients vector $\boldsymbol{\beta}_0$ by setting $\boldsymbol{\beta}_1 = (1.25, -0.95, 0.9, -1.1, 0.6)^T$. For each simulated data set, the response vector $\mathbf{y}$ was generated independently from the Poisson distribution with conditional mean vector $\exp(\mathbf{X}\boldsymbol{\beta}_0)$. The regularization parameter λ was selected by BIC (SIC performed similarly to BIC).

The PE is defined as $E[Y - \exp(\mathbf{X}^T\widehat{\boldsymbol{\beta}})]^2$, where $\widehat{\boldsymbol{\beta}}$ is the estimated coefficients vector by a method and $(\mathbf{X}^T, Y)$ is an independent test point. Lasso, SCAD and MCP had FN = 0 over 100 simulations. Table 3 summarizes the comparison results given by PE, L_2 loss, L_1 loss, deviance, #S, and FN.

We also examined the performance of non-concave penalized likelihood methods in high dimensional Poisson regression. The setting of this simulation is the same as above, except that $p = 500$ and 1000. The regularization parameter λ was selected by BIC and five-fold cross-validation (CV) based on prediction error. For both BIC and CV, Lasso had median FN = 0 with some nonzeros, and SCAD and MCP had FN = 0 over 100 simulations. Table

Table 2: Medians and robust standard deviations (in parentheses) of PE, L_2 loss, L_1 loss, deviance, #S, and FN over 100 simulations for all methods in logistic regression, where $p = 500$ and 1000

p	Measures	Lasso	SCAD	MCP	Oracle
500	PE	0.0149(0.015)	0.095(0.006)	0.096(0.007)	0.094(0.002)
	L_2 loss	4.158(0.574)	1.054(1.054)	1.160(0.985)	0.834(0.452)
	L_1 loss	11.540(0.841)	2.508(2.044)	2.481(2.292)	1.591(0.939)
	Deviance	113.84(43.76)	100.22(16.03)	102.96(15.36)	108.06(17.33)
	#S	41(20.39)	9(3.71)	6(1.48)	5(0)
	FN	0(0.74)	0(0)	0(0)	0(0)
1000	PE	0.163(0.010)	0.096(0.020)	0.096(0.007)	0.093(0.003)
	L_2 loss	4.753(0.333)	1.400(1.591)	1.010(1.000)	0.808(0.517)
	L_1 loss	11.759(0.801)	3.133(3.297)	2.322(2.145)	1.490(0.949)
	Deviance	152.19(49.36)	99.18(19.80)	103.25(16.99)	110.03(14.49)
	#S	28.5(24.83)	13(4.45)	7(2.22)	5(0)
	FN	1(0.74)	0(0)	0(0)	0(0)

Table 3: Medians and robust standard deviations (in parentheses) of PE, L_2 loss, L_1 loss, deviance, #S, and FN over 100 simulations for all methods in Poisson regression, where $p = 25$

Measures	Lasso	SCAD	MCP	Oracle
PE	7.195(2.428)	4.081(0.826)	4.012(0.791)	3.688(0.574)
L_2 loss	0.269(0.076)	0.141(0.045)	0.136(0.040)	0.111(0.035)
L_1 loss	0.606(0.215)	0.276(0.103)	0.271(0.094)	0.216(0.067)
Deviance	191.09(14.62)	186.73(12.72)	187.23(13.14)	187.72(15.28)
#S	9(2.22)	5(0.74)	5(0.74)	5(0)
FN	0(0)	0(0)	0(0)	0(0)

Table 4: Medians and robust standard deviations (in parentheses) of PE, L_2 loss, L_1 loss, deviance, #S, and FN over 100 simulations for all methods in Poisson regression by BIC and CV, where $p = 500$ and 1000

p	Method	Measures	Lasso	SCAD	MCP	Oracle
500	BIC	PE	26.989(11.339)	4.820(1.772)	4.672(1.593)	3.479(0.738)
		L_2 loss	0.790(0.206)	0.199(0.074)	0.178(0.076)	0.104(0.043)
		L_1 loss	2.446(0.638)	0.424(0.165)	0.371(0.161)	0.184(0.083)
		Deviance	202.65(22.23)	187.15(16.41)	189.66(17.24)	189.30(21.73)
		#S	29(5.93)	9(3.34)	7(2.22)	5(0)
		FN	0(0)	0(0)	0(0)	0(0)
	CV	PE	24.200(9.636)	4.542(1.554)	4.272(1.503)	3.479(0.738)
		L_2 loss	0.698(0.162)	0.168(0.065)	0.168(0.057)	0.104(0.043)
		L_1 loss	3.229(1.368)	0.495(0.201)	0.411(0.162)	0.184(0.083)
		Deviance	117.16(40.13)	166.58(21.76)	173.01(19.17)	189.30(21.73)
		#S	63.5(24.83)	18(10.75)	12.5(6.67)	5(0)
		FN	0(0)	0(0)	0(0)	0(0)
1000	BIC	PE	33.069(14.089)	5.523(2.027)	5.144(1.808)	3.676(0.772)
		L_2 loss	0.971(0.209)	0.210(0.094)	0.187(0.088)	0.108(0.047)
		L_1 loss	2.990(0.689)	0.485(0.232)	0.443(0.198)	0.197(0.090)
		Deviance	199.99(22.89)	180.34(13.07)	181.21(15.31)	187.98(17.22)
		#S	34(7.41)	11.5(4.08)	9(2.22)	5(0)
		FN	0(0)	0(0)	0(0)	0(0)
	CV	PE	31.701(16.571)	4.821(1.732)	4.700(1.702)	3.676(0.772)
		L_2 loss	0.889(0.201)	0.162(0.077)	0.162(0.064)	0.108(0.047)
		L_1 loss	4.297(1.646)	0.506(0.341)	0.454(0.239)	0.197(0.090)
		Deviance	92.89(44.51)	160.23(20.80)	169.34(23.44)	187.98(17.22)
		#S	83(40.77)	22(11.86)	14(7.04)	5(0)
		FN	0(0)	0(0)	0(0)	0(0)

Table 5: Classification errors in the neuroblastoma data set

	3-year EFS		Gender	
Method	# of genes	Test error	# of genes	Test error
Lasso	56	23/114	4	5/126
SCAD	10	18/114	2	4/126
MCP	7	23/114	1	12/126
SIS	5	19/114	6	4/126
ISIS	23	22/114	2	4/126

4 reports the comparison results given by PE, L_2 loss, L_1 loss, deviance, #S, and FN.

6.3 Real data analysis

In this example, we apply non-concave penalized likelihood methods to the neuroblastoma data set, which was studied by Oberthuer *et al.* (2006). This data set, obtained via the MicroArray Quality Control phase-II (MAQC-II) project, consists of gene expression profiles for 10,707 genes from 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004. The patients at diagnosis were aged from 0 to 296 months with a median age of 15 months. The study aimed to develop a gene expression-based classifier for neuroblastoma patients that can reliably predict courses of the disease.

We analyzed this data set for two binary responses: 3-year event-free survival (3-year EFS) and gender, where 3-year EFS indicates whether a patient survived 3 years after the diagnosis of neuroblastoma. There are 246 subjects with 101 females and 145 males, and 239 of them have the 3-year EFS information available (49 positives and 190 negatives). We applied Lasso, SCAD and MCP using the logistic regression model. Five-fold cross-validation was used to select the tuning parameter. For the 3-year EFS classification, we randomly selected 125 subjects (25 positives and 100 negatives) as the training set and the rest as the test set. For the gender classification, we randomly chose 120 subjects (50 females and 70 males) as the training set and the rest as the test set. Table 5 reports the classification results of all methods, as well as those of SIS and ISIS, which were extracted from Fan *et al.* (2009). Tables 6 and 7 list the selected genes by Lasso, SCAD and MCP for the 3-year EFS classification and gender classification, respectively.

7 Discussions

We have studied penalized likelihood methods for ultra-high dimensional variable selection. In the context of GLMs, we have shown that such methods have model selection consistency with oracle properties even for NP-dimensionality, for a class of non-concave penalized

Table 6: Selected genes for the 3-year EFS classification

Gene	Lasso	SCAD	MCP	Gene	Lasso	SCAD	MCP
A_24_P182182		x		Hs419768.1	x		
A_23_P144096	x			A_23_P313728	x		
A_23_P124514	x			A_23_P12884	x		
A_23_P502879	x			A_23_P130626	x		
A_23_P71319	x			A_23_P143958		x	
A_24_P73158	x	x		Hs155462.1		x	
A_24_P282251	x		x	A_23_P209394	x		
A_23_P125435	x			A_24_P100419	x		
A_23_P80491	x			Hs379382.1	x		
A_23_P77779	x			A_24_P271696	x		
A_23_P36076	x	x		Hs381187.1	x		
A_23_P35349	x			Hs265827.1	x		
A_23_P208030	x			Hs269914.3	x		
A_23_P72737	x		x	Hs36034.1	x		
A_23_P25194	x			A_23_P83751	x		
A_23_P200043	x			A_23_P96325	x		
A_23_P422809	x			A_23_P97553	x		
A_23_P110345	x		x	A_24_P232158	x		
A_23_P5131	x		x	A_23_P9836	x		
A_23_P11859	x			Hs170298.1	x		x
A_23_P7376	x			r60_a135	x		
A_23_P211985	x			A_23_P133956	x		
A_24_P365954		x		A_32_P27511	x		
A_23_P86975	x			A_23_P80626	x		
A_23_P89910	x			A_32_P158708	x		
A_24_P285055	x			A_23_P100764		x	
A_23_P68547	x			Hs407755.1	x		
A_23_P6252	x			Hs86643.1	x		
A_23_P386356	x			Hs422789.1	x	x	
A_24_P50228			x	A_23_P131866		x	
Hs37637.1	x			A_23_P147397	x		
Hs457415.1		x	x	A_23_P13852	x		

Table 7: Selected genes for the gender classification

Gene	Lasso	SCAD	MCP
A_23_P329835	x		
A_23_P259314	x		
A_23_P137238	x	x	x
A_24_P500584	x	x	

likelihood approaches. Our results are consistent with a known fact in the literature that concave penalties can reduce the bias problems of convex penalties. The convex function of L_1-penalty falls at the boundary of the class of penalty functions under consideration. We have used the coordinate optimization to find the solution paths and illustrated the performance of non-concave penalized likelihood methods with numerical studies. Our results show that the coordinate optimization works equally well and efficiently for producing the entire solution paths for concave penalties.

8 Proofs

8.1 Proof of Theorem 1

We will first derive the necessary condition. In view of (2), we have

$$\nabla \ell_n(\boldsymbol{\beta}) = n^{-1}\left[\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}\left(\boldsymbol{\theta}\right)\right] \quad \text{and} \quad \nabla^2 \ell_n(\boldsymbol{\beta}) = -n^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)\mathbf{X}, \tag{30}$$

where $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$. It follows from the classical optimization theory that if $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ is a local maximizer of the penalized likelihood (3), it satisfies the Karush-Kuhn-Tucker (KKT) conditions, i.e., there exists some $\mathbf{v} = (v_1, \cdots, v_p)^T \in \mathbf{R}^p$ such that

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) - n\lambda_n\mathbf{v} = \mathbf{0}, \tag{31}$$

where $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, $v_j = \bar{\rho}(\widehat{\beta}_j)$ for $\widehat{\beta}_j \neq 0$, and $v_j \in [-\rho'(0+), \rho'(0+)]$ for $\widehat{\beta}_j = 0$. Let $\mathcal{S} = \text{supp}(\widehat{\boldsymbol{\beta}})$. Note that $\widehat{\boldsymbol{\beta}}$ is also a local maximizer of (3) constrained on the $\|\widehat{\boldsymbol{\beta}}\|_0$-dimensional subspace $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbf{R}^p : \boldsymbol{\beta}_c = \mathbf{0}\}$ of $\mathbf{R}^p$, where $\boldsymbol{\beta}_c$ denotes the subvector of $\boldsymbol{\beta}$ formed by components in $\mathcal{S}^c$, the complement of $\mathcal{S}$. It follows from the second order condition that

$$\lambda_{\min}\left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{X}_1\right] \geq n\lambda_n\kappa(\rho;\widehat{\boldsymbol{\beta}}_1), \tag{32}$$

where $\kappa(\rho;\widehat{\boldsymbol{\beta}}_1)$ is given by (6). It is easy to see that equation (31) can be equivalently written as

$$\mathbf{X}_1^T\mathbf{y} - \mathbf{X}_1^T\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}) - n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\beta}}_1) = \mathbf{0}, \tag{33}$$

$$\|\mathbf{z}\|_\infty \leq \rho'(0+), \tag{34}$$

where $\mathbf{z} = (n\lambda_n)^{-1}\mathbf{X}_2^T[\mathbf{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})]$ and $\mathbf{X}_2$ denotes the submatrix of $\mathbf{X}$ formed by columns in $\mathcal{S}^c$.

We now prove the sufficient condition. We first constrain the penalized likelihood (3) on the $\|\widehat{\boldsymbol{\beta}}\|_0$-dimensional subspace $\mathcal{B}$ of $\mathbf{R}^p$. It follows from condition (9) that $Q_n(\boldsymbol{\beta})$ is strictly concave in a ball $\mathcal{N}_0$ in the subspace $\mathcal{B}$ centered at $\widehat{\boldsymbol{\beta}}$. This along with equation (7) immediately entails that $\widehat{\boldsymbol{\beta}}$, as a critical point of $Q_n(\boldsymbol{\beta})$ in $\mathcal{B}$, is the unique maximizer of $Q_n(\boldsymbol{\beta})$ in the neighborhood $\mathcal{N}_0$.

It remains to prove that the sparse vector $\widehat{\boldsymbol{\beta}}$ is indeed a strict local maximizer of $Q_n(\boldsymbol{\beta})$ on the space $\mathbf{R}^p$. To show this, take a sufficiently small ball $\mathcal{N}_1$ in $\mathbf{R}^p$ centered at $\widehat{\boldsymbol{\beta}}$ such that $\mathcal{N}_1 \cap \mathcal{B} \subset \mathcal{N}_0$. We then need to show that $Q_n(\widehat{\boldsymbol{\beta}}) > Q_n(\boldsymbol{\gamma}_1)$ for any $\boldsymbol{\gamma}_1 \in \mathcal{N}_1 \setminus \mathcal{N}_0$. Let $\boldsymbol{\gamma}_2$ be the projection of $\boldsymbol{\gamma}_1$ onto the subspace $\mathcal{B}$. Then we have $\boldsymbol{\gamma}_2 \in \mathcal{N}_0$, which entails that $Q_n(\widehat{\boldsymbol{\beta}}) > Q_n(\boldsymbol{\gamma}_2)$ if $\boldsymbol{\gamma}_2 \neq \widehat{\boldsymbol{\beta}}$, since $\widehat{\boldsymbol{\beta}}$ is the strict maximizer of $Q_n(\boldsymbol{\beta})$ in $\mathcal{N}_0$. Thus, it suffices to show that $Q_n(\boldsymbol{\gamma}_2) > Q_n(\boldsymbol{\gamma}_1)$.

By the mean-value theorem, we have

$$Q_n(\boldsymbol{\gamma}_1) - Q_n(\boldsymbol{\gamma}_2) = \nabla^T Q_n(\boldsymbol{\gamma}_0)(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2), \tag{35}$$

where $\boldsymbol{\gamma}_0$ lies on the line segment joining $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_1$. Note that the components of $\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2$ are zero for the indices in $\mathcal{S}$ and the sign of $\gamma_{0,j}$ is the same as that of $\gamma_{1,j}$ for $j \notin \mathcal{S}$, where $\gamma_{0,j}$ and $\gamma_{1,j}$ are the j-th components of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$, respectively. Therefore, the right hand side of (35) can be expressed as

$$\left\{n^{-1}\mathbf{X}_2^T\left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\gamma}_0)\right]\right\}^T \boldsymbol{\gamma}_{1,2} - \lambda_n \sum_{j \notin \mathcal{S}} \rho'(|\gamma_{0,j}|)|\gamma_{1,j}|, \tag{36}$$

where $\boldsymbol{\gamma}_{1,2}$ is a subvector of $\boldsymbol{\gamma}_1$ formed by the components in $\mathcal{S}^c$. By $\boldsymbol{\gamma}_1 \in \mathcal{N}_1 \setminus \mathcal{N}_0$, we have $\boldsymbol{\gamma}_{1,2} \neq \mathbf{0}$.

It follows from the concavity of ρ in Condition 1 that $\rho'(t)$ is decreasing in $t \in [0, \infty)$. By condition (8) and the continuity of $\rho'(t)$ and $b'(\theta)$, there exists some $\delta > 0$ such that for any $\boldsymbol{\beta}$ in a ball in $\mathbf{R}^p$ centered at $\widehat{\boldsymbol{\beta}}$ with radius δ,

$$\|(n\lambda_n)^{-1}\mathbf{X}_2^T\left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})\right]\|_\infty < \rho'(\delta). \tag{37}$$

We further shrink the radius of the ball $\mathcal{N}_1$ to less than δ so that $|\gamma_{0,j}| \leq |\gamma_{1,j}| < \delta$ for $j \notin \mathcal{S}$ and (37) holds for any $\boldsymbol{\beta} \in \mathcal{N}_1$. Since $\boldsymbol{\gamma}_0 \in \mathcal{N}_1$, it follows from (37) that the term (36) is strictly less than

$$\lambda_n \rho'(\delta)\|\boldsymbol{\gamma}_{1,2}\|_1 - \lambda_n \rho'(\delta)\|\boldsymbol{\gamma}_{1,2}\|_1 = 0,$$

where the monotonicity of $\rho'(\cdot)$ was used in the second term. Thus we conclude that $Q_n(\gamma_1) < Q_n(\gamma_2)$. This completes the proof.

8.2 Proof of Proposition 1

Let $\partial\mathcal{L}_c = \{\boldsymbol{\beta} \in \mathbf{R}^p : \ell_n(\boldsymbol{\beta}) = c\}$ be the level set. By the concavity of $\ell_n(\boldsymbol{\beta})$, we can easily show that for $c < \ell_n(\mathbf{0})$, $\mathcal{L}_c$ is a closed convex set with $\boldsymbol{\beta}_*$ and $\mathbf{0}$ being its interior points and the level set $\partial\mathcal{L}_c$ is its boundary. We now show that the global maximizer of the penalized likelihood $Q_n(\boldsymbol{\beta})$ belongs to $\mathcal{L}_c$.

For any $\boldsymbol{\gamma} \in \partial\mathcal{L}_c$, let $\Gamma\boldsymbol{\gamma} = \{t\boldsymbol{\gamma} : t \in (1, \infty)\}$ be a ray. By the convexity of $\mathcal{L}_c$, we have $\{t\boldsymbol{\gamma} : t \in [0, 1]\} \subset \mathcal{L}_c$ for $\boldsymbol{\gamma} \in \partial\mathcal{L}_c$, which implies that

$$\bigcup_{\boldsymbol{\gamma}\in\partial\mathcal{L}_c} \Gamma\boldsymbol{\gamma} = \mathbf{R}^p \setminus \mathcal{L}_c.$$

Thus to show that the global maximizer of $Q_n(\boldsymbol{\beta})$ belongs to $\mathcal{L}_c$, it suffices to prove $Q_n(t\boldsymbol{\gamma}) < Q_n(\boldsymbol{\gamma})$ for any $t \in (1, \infty)$ and $\boldsymbol{\gamma} \in \partial\mathcal{L}_c$. This follows easily from the definition of $Q_n(\boldsymbol{\beta})$, $\ell_n(t\boldsymbol{\gamma}) < c = \ell_n(\boldsymbol{\gamma})$, and $\sum_{j=1}^p p_{\lambda_n}(t|\gamma_j|) \geq \sum_{j=1}^p p_{\lambda_n}(|\gamma_j|)$, where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^T$.

It remains to prove that the local maximizer of $Q_n(\boldsymbol{\beta})$ in $\mathcal{L}_c$ must be a global maximizer. This is entailed by the concavity of $Q_n(\boldsymbol{\beta})$ on $\mathcal{L}_c$, which is ensured by condition (11). This concludes the proof.

8.3 Proof of Proposition 2

Since $c < \ell_n(\mathbf{0})$, from the proof of Proposition 1 we know that the global maximizer of the penalized likelihood $Q_n(\boldsymbol{\beta})$ belongs to $\mathcal{L}_c$. Note that by assumption, the SCAD penalized likelihood estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T \in \mathcal{L}_c$ and $\min_{j=1}^p |\widehat{\beta}_j| > a\lambda_n$. It follows from (3) and (4) that $\widehat{\boldsymbol{\beta}}$ is a critical point of $\ell_n(\boldsymbol{\beta})$ and thus $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_*$ by the strict concavity of $\ell_n(\boldsymbol{\beta})$. It remains to prove that $\boldsymbol{\beta}_*$ is the maximizer of $Q_n(\boldsymbol{\beta})$ on $\mathcal{L}_c$.

The key idea is to use a first order Taylor expansion of $\ell_n(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_*$ and retain the Lagrange remainder term. This along with $\nabla\ell_n(\boldsymbol{\beta}_*) = \mathbf{0}$ and $\min_{\boldsymbol{\beta}\in\mathcal{L}_c} \lambda_{\min}[n^{-1}\mathbf{X}^T\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}] \geq c_0$ gives for any $\boldsymbol{\beta} \in \mathcal{L}_c$,

$$Q_n(\boldsymbol{\beta}) \leq \widetilde{Q}_n(\boldsymbol{\beta}) \equiv \ell_n(\boldsymbol{\beta}_*) - \frac{c_0}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2^2 - \sum_{j=1}^p p_{\lambda_n}(|\beta_j|),$$

since $\boldsymbol{\beta}_*$ is in the convex set $\mathcal{L}_c$. Thus if $\boldsymbol{\beta}_*$ is the global maximizer of $\widetilde{Q}_n(\boldsymbol{\beta})$ on $\mathbf{R}^p$, then we have for any $\boldsymbol{\beta} \in \mathcal{L}_c$,

$$Q_n(\boldsymbol{\beta}) \leq \widetilde{Q}_n(\boldsymbol{\beta}) \leq \widetilde{Q}_n(\boldsymbol{\beta}_*) = Q_n(\boldsymbol{\beta}_*).$$

This entails that $\boldsymbol{\beta}_*$ is the global maximizer of $Q_n(\boldsymbol{\beta})$.

To maximize $\widetilde{Q}_n(\boldsymbol{\beta})$, we only need to maximize it componentwise. Let $\boldsymbol{\beta}_* = (\beta_{*,1}, \cdots, \beta_{*,p})^T$. Then it remains to show that for each $j = 1, \cdots, p$, $\beta_{*,j}$ is the global minimizer of the univariate SCAD penalized least squares problem

$$\min_{\beta\in\mathbf{R}} g_j(\beta) = \min_{\beta\in\mathbf{R}} \left\{ \frac{c_0}{2}(\beta - \beta_{*,j})^2 + p_{\lambda_n}(|\beta|) \right\}. \tag{38}$$

This can easily been shown from the analytical solution to (38). For the sake of completeness, we give a simple proof here.

Recall that we have shown that $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_*$. In view of (38) and $|\beta_{*,j}| > a\lambda_n$, for any $|\beta| > a\lambda_n$ with $\beta \neq \beta_{*,j}$, we have

$$g_j(\beta) > p_{\lambda_n}(|\beta|) = p_{\lambda_n}(a\lambda_n) = g_j(|\beta_{*,j}|),$$

where we used the fact that $p_{\lambda_n}(\cdot)$ is constant on $[a\lambda_n, \infty)$. Thus, it suffices to prove $g_j(\beta) > p_{\lambda_n}(a\lambda_n)$ on the interval $|\beta| \leq a\lambda_n$. For such a β, we have $p_{\lambda_n}(a\lambda_n) - p_{\lambda_n}(|\beta|) \leq \lambda_n(a\lambda_n - |\beta|)$. Thus we need to show that

$$\min_{z \in [0, a\lambda_n]} \left\{ \frac{c_0}{2} \left(|\beta_{*,j}| - a\lambda_n + z \right)^2 - \lambda_n z \right\} > 0,$$

which always holds as long as $|\beta_{*,j}| > (a + \frac{1}{2c_0})\lambda_n$ and thus completes the proof.

8.4 Proof of Proposition 3

Let $\mathcal{A}$ be any s-dimensional coordinate subspace different from $\mathcal{A}_1 = \{(\beta_1, \cdots, \beta_p)^T \in \mathbf{R}^p : \beta_j = 0 \text{ for } j \notin \text{supp}(\widehat{\boldsymbol{\beta}})\}$. Clearly $\mathcal{A}_1 \oplus \mathcal{A}$ is a d-dimensional coordinate subspace with $d \leq 2s$. Then part a) follows easily from the assumptions and Proposition 1. Part b) is an easy consequence of Proposition 2 in view of the assumptions and the fact that

$$\max_{t \in [0, \infty)} p_{\lambda_n}(t) = p_{\lambda_n}(a\lambda_n) = \frac{(a+1)\lambda_n^2}{2}$$

for the SCAD penalty p_λ given by (4).

8.5 Proof of Proposition 4

Part a) follows easily from a simple application of Hoeffding's inequality (Hoeffding, 1963), since $a_1 Y_1, \cdots, a_n Y_n$ are n independent bounded random variables, where $\mathbf{a} = (a_1, \cdots, a_n)^T$. We now prove part b). In view of condition (20), $a_i Y_i - a_i b'(\theta_{0,i})$ are n independent random variables with mean zero and satisfy

$$\begin{aligned} E\left| a_i Y_i - a_i b'\left(\theta_{0,i}\right) \right|^m &= |a_i|^m E\left| Y_i - b'\left(\theta_{0,i}\right) \right|^m \leq |a_i|^m m! M^{m-2} \frac{v_0}{2} \\ &\leq \frac{m!}{2} \left(\|\mathbf{a}\|_\infty M \right)^{m-2} a_i^2 v_0, \quad m \geq 2. \end{aligned}$$

Thus an application of Bernstein's inequality (see, e.g., Bennett, 1962 or van der Vaart and Wellner, 1996) yields

$$\begin{aligned} P\left(\left| \mathbf{a}^T \mathbf{Y} - \mathbf{a}^T \boldsymbol{\mu}(\boldsymbol{\theta}_0) \right| > \varepsilon \right) &\leq 2 \exp\left[-\frac{1}{2} \frac{\varepsilon^2}{\sum_{i=1}^n a_i^2 v_0 + \|\mathbf{a}\|_\infty M\varepsilon} \right] \\ &= 2 \exp\left[-\frac{1}{2} \frac{\varepsilon^2}{\|\mathbf{a}\|_2^2 v_0 + \|\mathbf{a}\|_\infty M\varepsilon} \right], \end{aligned}$$

which concludes the proof.

8.6 Proof of Theorem 2

We break the whole proof into several steps. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively be the submatrices of $\mathbf{X}$ formed by columns in $\mathfrak{M}_0 = \text{supp}(\boldsymbol{\beta}_0)$ and its complement $\mathfrak{M}_0^c$, and $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0$. Let $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}(\boldsymbol{\theta}_0)$. Consider events

$$\mathcal{E}_1 = \left\{ \left\| \boldsymbol{\xi}_{\mathfrak{M}_0} \right\|_\infty \le c_1^{-1/2}\sqrt{n \log n} \right\} \quad \text{and} \quad \mathcal{E}_2 = \left\{ \left\| \boldsymbol{\xi}_{\mathfrak{M}_0^c} \right\|_\infty \le u_n\sqrt{n} \right\},$$

where $u_n = c_1^{-1/2} n^{1/2-\alpha} (\log n)^{1/2}$ is a diverging sequence and $\mathbf{v}_A$ denotes a subvector of $\mathbf{v}$ consisting of elements in A. Since $\|\mathbf{x}_j\|_2 = \sqrt{n}$, it follows from Bonferroni's inequality and (22) that

$$\begin{aligned} & P\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \\ & \ge 1 - \sum_{j \in \mathfrak{M}_0} P\left(|\xi_j| > c_1^{-1/2}\sqrt{n \log n} \right) - \sum_{j \in \mathfrak{M}_0^c} P\left(|\xi_j| > u_n\sqrt{n} \right) \\ & \ge 1 - 2\left[s n^{-1} + (p-s)\, e^{-c_1 u_n^2} \right] \\ & = 1 - 2[s n^{-1} + (p-s) e^{-n^{1-2\alpha} \log n}], \end{aligned} \tag{39}$$

where $s = \|\boldsymbol{\beta}_0\|_0$ and $u_n \le \sqrt{n} / \max_{j=1}^p \|\mathbf{x}_j\|_\infty$ for unbounded responses, which is guaranteed for sufficiently large n by Condition 3. Under the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we will show that there exists a solution $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ to (7)–(9) with $\text{sgn}(\widehat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta}_0)$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty = O(n^{-\gamma} \log n)$, where the function sgn is applied componentwise.

Step 1: Existence of a solution to equation (7). We first prove that for sufficiently large n, equation (7) has a solution $\widehat{\boldsymbol{\beta}}_1$ inside the hypercube

$$\mathcal{N} = \left\{ \boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_\infty = n^{-\gamma} \log n \right\}.$$

For any $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_s)^T \in \mathcal{N}$, since $d_n \ge n^{-\gamma} \log n$, we have

$$\min_{j=1}^s |\delta_j| \ge \min_{j \in \mathfrak{M}_0} |\beta_{0,j}| - d_n = d_n \tag{40}$$

and $\text{sgn}(\boldsymbol{\delta}) = \text{sgn}(\boldsymbol{\beta}_1)$. Let $\boldsymbol{\eta} = n\lambda_n \bar{\rho}(\boldsymbol{\delta})$. Using the monotonicity condition of $\rho'(t)$, by (40) we have

$$\|\boldsymbol{\eta}\|_\infty \le n\lambda_n \rho'(d_n),$$

which along with the definition of $\mathcal{E}_1$ entails

$$\|\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta}\|_\infty \le c_1^{-1/2}\sqrt{n \log n} + n\lambda_n \rho'(d_n). \tag{41}$$

Define vector-valued functions

$$\boldsymbol{\gamma}(\boldsymbol{\delta}) = (\gamma_1(\boldsymbol{\delta}), \cdots, \gamma_p(\boldsymbol{\delta}))^T = \mathbf{X}^T\boldsymbol{\mu}(\mathbf{X}_1\boldsymbol{\delta}), \quad \boldsymbol{\delta} \in \mathbf{R}^s$$

and

$$\boldsymbol{\Psi}(\boldsymbol{\delta}) = \boldsymbol{\gamma}_{\mathfrak{M}_0}(\boldsymbol{\delta}) - \boldsymbol{\gamma}_{\mathfrak{M}_0}(\boldsymbol{\beta}_1) - (\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta}), \quad \boldsymbol{\delta} \in \mathbf{R}^s.$$

Then, equation (7) is equivalent to $\boldsymbol{\Psi}(\boldsymbol{\delta}) = \mathbf{0}$. We need to show that the latter has a solution inside the hypercube $\mathcal{N}$. To this end, we represent $\boldsymbol{\gamma}_{\mathfrak{M}_0}(\boldsymbol{\delta})$ by using a second order Taylor expansion around $\boldsymbol{\beta}_1$ with the Lagrange remainder term componentwise and obtain

$$\boldsymbol{\gamma}_{\mathfrak{M}_0}(\boldsymbol{\delta}) = \boldsymbol{\gamma}_{\mathfrak{M}_0}(\boldsymbol{\beta}_1) + \mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1 (\boldsymbol{\delta} - \boldsymbol{\beta}_1) + \mathbf{r}, \tag{42}$$

where $\mathbf{r} = (r_1, \cdots, r_s)^T$ and for each $j = 1, \cdots, s$,

$$r_j = \frac{1}{2} \left(\boldsymbol{\delta} - \boldsymbol{\beta}_1\right)^T \nabla^2 \gamma_j(\boldsymbol{\delta}_j) \left(\boldsymbol{\delta} - \boldsymbol{\beta}_1\right)$$

with $\boldsymbol{\delta}_j$ some s-vector lying on the line segment joining $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_1$. By (17), we have

$$\begin{aligned} \|\mathbf{r}\|_\infty &\leq \max_{\boldsymbol{\delta}_0 \in \mathcal{N}} \max_{j=1}^{s} \frac{1}{2} \lambda_{\max} \left[\mathbf{X}_1^T \operatorname{diag}\left\{|\mathbf{x}_j| \circ \left|\boldsymbol{\mu}''\left(\mathbf{X}_1 \boldsymbol{\delta}_0\right)\right|\right\} \mathbf{X}_1\right] \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_2^2 \\ &= O\left[s n^{1-2\gamma} (\log n)^2\right]. \end{aligned} \tag{43}$$

Let

$$\overline{\boldsymbol{\Psi}}(\boldsymbol{\delta}) \equiv \left[\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1\right]^{-1} \boldsymbol{\Psi}(\boldsymbol{\delta}) = \boldsymbol{\delta} - \boldsymbol{\beta}_1 + \mathbf{u}, \tag{44}$$

where $\mathbf{u} = -[\mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1]^{-1} (\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta} - \mathbf{r})$. It follows from (41), (43), and (15) in Condition 2 that for any $\boldsymbol{\delta} \in \mathcal{N}$,

$$\begin{aligned} \|\mathbf{u}\|_\infty &\leq \left\| \left[\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1\right]^{-1} \right\|_\infty \left(\|\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta}\|_\infty + \|\mathbf{r}\|_\infty\right) \\ &= O\left[b_s n^{-1/2} \sqrt{\log n} + b_s \lambda_n \rho'(d_n) + b_s s n^{-2\gamma} (\log n)^2\right]. \end{aligned} \tag{45}$$

By Condition 3, the first and third terms are of order $o(n^{-\gamma} \log n)$ and so is the second term by (18). This shows that

$$\|\mathbf{u}\|_\infty = o(n^{-\gamma} \log n).$$

By (44), for sufficiently large n, if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1)_j = n^{-\gamma} \sqrt{\log n}$, we have

$$\overline{\Psi}_j(\boldsymbol{\delta}) \geq n^{-\gamma} \sqrt{\log n} - \|\mathbf{u}\|_\infty \geq 0, \tag{46}$$

and if $(\boldsymbol{\delta} - \boldsymbol{\beta}_1)_j = -n^{-\gamma} \sqrt{\log n}$, we have

$$\overline{\Psi}_j(\boldsymbol{\delta}) \leq -n^{-\gamma} \sqrt{\log n} + \|\mathbf{u}\|_\infty \leq 0, \tag{47}$$

where $\overline{\boldsymbol{\Psi}}(\boldsymbol{\delta}) = (\overline{\Psi}_1(\boldsymbol{\delta}), \cdots, \overline{\Psi}_s(\boldsymbol{\delta}))^T$. By the continuity of the vector-valued function $\overline{\boldsymbol{\Psi}}(\boldsymbol{\delta})$, (46) and (47), an application of Miranda's existence theorem (see, e.g., Vrahatis, 1989) shows that equation $\overline{\boldsymbol{\Psi}}(\boldsymbol{\delta}) = \mathbf{0}$ has a solution $\widehat{\boldsymbol{\beta}}_1$ in $\mathcal{N}$. Clearly $\widehat{\boldsymbol{\beta}}_1$ also solves equation $\boldsymbol{\Psi}(\boldsymbol{\delta}) = \mathbf{0}$ in view of (44). Thus we have shown that equation (7) indeed has a solution $\widehat{\boldsymbol{\beta}}_1$ in $\mathcal{N}$.

Step 2: Verification of condition (8). Let $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ with $\widehat{\boldsymbol{\beta}}_{\mathfrak{M}_0} = \widehat{\boldsymbol{\beta}}_1 \in \mathcal{N}$ a solution to equation (7) and $\widehat{\boldsymbol{\beta}}_{\mathfrak{M}_0^c} = \mathbf{0}$, and $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. We now show that $\widehat{\boldsymbol{\beta}}$ satisfies inequality (8) for λ_n given by (18). Note that

$$\begin{aligned} \mathbf{z} &= (n\lambda_n)^{-1}\left\{\left[\mathbf{X}_2^T\mathbf{y} - \mathbf{X}_2^T\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right] - \left[\mathbf{X}_2^T\boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{X}_2^T\boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right)\right]\right\} \\ &= (n\lambda_n)^{-1}\left\{\boldsymbol{\xi}_{\mathfrak{M}_0^c} - \left[\boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\widehat{\boldsymbol{\beta}}_1) - \boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\boldsymbol{\beta}_1)\right]\right\}. \end{aligned} \tag{48}$$

On the event $\mathcal{E}_2$, the L_∞ norm of the first term is bounded by $O(n^{-1/2}u_n\lambda_n^{-1}) = o(1)$ by the condition on λ_n. It remains to bound the second term of (48).

A Taylor expansion of $\boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\boldsymbol{\delta})$ around $\boldsymbol{\beta}_1$ componentwise gives

$$\boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\boldsymbol{\beta}_1) + \mathbf{X}_2^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) + \mathbf{w}, \tag{49}$$

where $\mathbf{w} = (w_{s+1}, \cdots, w_p)^T$ with $w_j = \frac{1}{2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)^T\nabla^2\gamma_j(\boldsymbol{\delta}_j)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)$ and $\boldsymbol{\delta}_j$ some s-vector lying on the line segment joining $\widehat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_1$. By (17) in Condition 2 and $\widehat{\boldsymbol{\beta}}_1 \in \mathcal{N}$, arguing similarly to (43), we have

$$\|\mathbf{w}\|_\infty = O\left[sn^{1-2\gamma}(\log n)^2\right]. \tag{50}$$

Since $\widehat{\boldsymbol{\beta}}_1$ solves equation $\overline{\boldsymbol{\Psi}}(\boldsymbol{\delta}) = \mathbf{0}$ in (44), we have

$$\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 = \left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\right]^{-1}\left(\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta} - \mathbf{r}\right). \tag{51}$$

It follows from (15) and (16) in Condition 2, (41), (43), and (48)–(51) that

$$\begin{aligned} \|\mathbf{z}\|_\infty &\le o(1) + (n\lambda_n)^{-1}\left\|\boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\widehat{\boldsymbol{\beta}}_1) - \boldsymbol{\gamma}_{\mathfrak{M}_0^c}(\boldsymbol{\beta}_1)\right\|_\infty \\ &\le o(1) + (n\lambda_n)^{-1}\left\|\mathbf{X}_2^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\right]^{-1}\right\|_\infty \\ &\quad \cdot\left(\left\|\boldsymbol{\xi}_{\mathfrak{M}_0} - \boldsymbol{\eta}\right\|_\infty + \|\mathbf{r}\|_\infty\right) + (n\lambda_n)^{-1}\|\mathbf{w}\|_\infty \\ &\le o(1) + (n\lambda_n)^{-1}O\left\{n^{\alpha_1}\left[\sqrt{n\log n} + sn^{1-2\gamma}(\log n)^2\right] + sn^{1-2\gamma}(\log n)^2\right\} \\ &\quad + \left\|\mathbf{X}_2^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\left[\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\mathbf{X}_1\right]^{-1}\right\|_\infty \rho'(d_n). \end{aligned}$$

The second term is of order $O(\lambda_n^{-1}n^{-\alpha}(\log n)^2) = o(1)$ by (18). Using (16), we have

$$\|\mathbf{z}\|_\infty \le C\rho'(0+) + o(1) < \rho'(0+)$$

for sufficiently large n.

Finally, note that condition (9) for sufficiently large n is guaranteed by $\lambda_n\kappa_0 = o(\tau_0)$ in Condition 3. Therefore, by Theorem 1, we have shown that $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ is a strict local maximizer of the non-concave penalized likelihood $Q_n(\boldsymbol{\beta})$ (3) with $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty = O(n^{-\gamma}\log n)$ and $\widehat{\boldsymbol{\beta}}_2 = 0$ under the event $\mathcal{E}_1 \cap \mathcal{E}_2$. These along with (39) prove parts a) and b). This completes the proof.

8.7 Proof of Theorem 3

We continue to adopt the notation in the proof of Theorem 2. To prove the conclusions, it suffices to show that under the given regularity conditions, there exists a strict local maximizer $\widehat{\boldsymbol{\beta}}$ of the penalized likelihood $Q_n(\boldsymbol{\beta})$ in (3) such that 1) $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$ (i.e., sparsity), and 2) $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_2 = O_P(\sqrt{s/n})$ (i.e., $\sqrt{s/n}$-consistency).

Step 1: Consistency in the s-dimensional subspace. We first constrain $Q_n(\boldsymbol{\beta})$ on the s-dimensional subspace $\{\boldsymbol{\beta} \in \mathbf{R}^p : \boldsymbol{\beta}_{\mathfrak{M}_0^c} = \mathbf{0}\}$ of $\mathbf{R}^p$. This constrained penalized likelihood is given by

$$\overline{Q}_n(\boldsymbol{\delta}) = \overline{\ell}_n(\boldsymbol{\delta}) - \sum_{j=1}^{s} p_{\lambda_n}(|\delta_j|), \tag{52}$$

where $\overline{\ell}_n(\boldsymbol{\delta}) = n^{-1}[\mathbf{y}^T\mathbf{X}_1\boldsymbol{\delta} - \mathbf{1}^T\mathbf{b}(\mathbf{X}_1\boldsymbol{\delta})]$ and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_s)^T$. We now show that there exists a strict local maximizer $\widehat{\boldsymbol{\beta}}_1$ of $\overline{Q}_n(\boldsymbol{\delta})$ such that $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_2 = O_P(\sqrt{s/n})$. To this end, we define an event

$$H_n = \left\{ \overline{Q}_n(\boldsymbol{\beta}_1) > \max_{\boldsymbol{\delta} \in \partial N_\tau} \overline{Q}_n(\boldsymbol{\delta}) \right\},$$

where ∂N_τ denotes the boundary of the closed set $N_\tau = \{\boldsymbol{\delta} \in \mathbf{R}^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_2 \le \sqrt{s/n}\tau\}$ and $\tau \in (0, \infty)$. Clearly, on the event H_n, there exists a local maximizer $\widehat{\boldsymbol{\beta}}_1$ of $\overline{Q}_n(\boldsymbol{\delta})$ in N_τ. Thus, we need only to show that $P(H_n)$ is close to 1 as $n \to \infty$ when τ is large. To this end, we need to analyze the function $\overline{Q}_n$ on the boundary ∂N_τ.

Let n be sufficiently large such that $\sqrt{s/n}\tau \le d_n$ since $d_n \gg \sqrt{s/n}$ by Condition 5. It is easy to see that $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_s)^T \in N_\tau$ entails $\mathrm{sgn}(\boldsymbol{\delta}) = \mathrm{sgn}(\boldsymbol{\beta}_1)$, $\|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_\infty \le d_n$, and $\min_j |\delta_j| \ge d_n$. By Taylor's theorem, we have for any $\boldsymbol{\delta} \in N_\tau$,

$$\overline{Q}_n(\boldsymbol{\delta}) - \overline{Q}_n(\boldsymbol{\beta}_1) = (\boldsymbol{\delta} - \boldsymbol{\beta}_1)^T\mathbf{v} - \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\beta}_1)^T\mathbf{D}(\boldsymbol{\delta} - \boldsymbol{\beta}_1), \tag{53}$$

where $\mathbf{v} = n^{-1}\mathbf{X}_1^T[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)] - \bar{p}_{\lambda_n}(\boldsymbol{\beta}_1)$, $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}_1\boldsymbol{\beta}_1$,

$$\mathbf{D} = n^{-1}\mathbf{X}_1^T\boldsymbol{\Sigma}\left(\boldsymbol{\theta}^*\right)\mathbf{X}_1 + \mathrm{diag}\left\{p''_{\lambda_n}(|\boldsymbol{\beta}^*|)\right\},$$

$\boldsymbol{\theta}^* = \mathbf{X}_1\boldsymbol{\beta}^*$, and $\boldsymbol{\beta}^*$ lies on the line segment joining $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_1$. More generally, when the second derivative of the penalty function p_λ does not necessarily exist, it is easy to show that the second part of the matrix $\mathbf{D}$ can be replaced by a diagonal matrix with maximum absolute element bounded by $\lambda_n\kappa_0$. Recall that

$$\mathcal{N}_0 = \{\mathbf{b} \in \mathbf{R}^s : \|\mathbf{b} - \boldsymbol{\beta}_1\|_\infty \le d_n\}$$

and $\kappa_0 = \max_{\mathbf{b} \in \mathcal{N}_0} \kappa(\rho; \mathbf{b})$, where $\kappa(\rho; \mathbf{b})$ is given by (6). For any $\boldsymbol{\delta} \in \partial N_\tau$, we have $\|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_2 = \sqrt{s/n}\tau$ and $\boldsymbol{\beta}^* \in \mathcal{N}_0$. Then for sufficiently large n, by (26) and $\lambda_n\kappa_0 = o(1)$ in Conditions 4 and 5 we have

$$\lambda_{\min}(\mathbf{D}) \ge c - \lambda_n\kappa_0 \ge \frac{c}{2}.$$

Thus by (53), we have

$$\max_{\boldsymbol{\delta}\in\partial N_\tau} \overline{Q}_n(\boldsymbol{\delta}) - \overline{Q}_n(\boldsymbol{\beta}_1) \le \sqrt{s/n}\tau \left(\|\mathbf{v}\|_2 - c\sqrt{s/n}\tau/4 \right),$$

which along with Markov's inequality entails that

$$P(H_n) \ge P\left(\|\mathbf{v}\|_2^2 < \frac{c^2 s\tau^2}{16n} \right) \ge 1 - \frac{16nE\|\mathbf{v}\|_2^2}{c^2 s\tau^2}.$$

It follows from $E\mathbf{y} = \boldsymbol{\mu}(\boldsymbol{\theta}_0)$, $\mathrm{cov}(\mathbf{y}) = \phi\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$, and Conditions 4 and 5 that

$$\begin{aligned} E\|\mathbf{v}\|_2^2 &= n^{-2} E \left\| \mathbf{X}_1^T \left[\mathbf{y} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right) \right] \right\|_2^2 + \|\bar{p}_{\lambda_n}(\boldsymbol{\beta}_1)\|_2^2 \\ &\le n^{-2}\phi \mathrm{tr}\left[\mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1 \right] + s p'_{\lambda_n}(d_n)^2 = O(sn^{-1}), \end{aligned}$$

since $p'_{\lambda_n}(t)$ is decreasing in $t \in [0, \infty)$. Hence, we have

$$P(H_n) \ge 1 - O(\tau^{-2}).$$

This proves $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_2 = O_P(\sqrt{s/n})$.

Step 2: Sparsity. Let $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ with $\widehat{\boldsymbol{\beta}}_{\mathfrak{M}_0} = \widehat{\boldsymbol{\beta}}_1 \in N_\tau \subset \mathcal{N}_0$ a strict local maximizer of $\overline{Q}_n(\boldsymbol{\delta})$ and $\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_{\mathfrak{M}_0^c} = \mathbf{0}$, and $\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. It remains to prove that the vector $\widehat{\boldsymbol{\beta}}$ is indeed a strict local maximizer of $Q_n(\boldsymbol{\beta})$ on the space $\mathbf{R}^p$. From the proof of Theorem 1, we see that it suffices to check condition (8). The idea is the same as that in Step 2 of the proof of Theorem 2. Let $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_p)^T = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}(\boldsymbol{\theta}_0)$ and consider the event

$$\mathcal{E}_2 = \left\{ \left\| \boldsymbol{\xi}_{\mathfrak{M}_0^c} \right\|_\infty \le u_n\sqrt{n} \right\},$$

where $u_n = c_1^{-1/2} n^{\alpha/2}\sqrt{\log n}$. We have shown in the proof of Theorem 2 that

$$P(\mathcal{E}_2) \ge 1 - (p-s)\varphi(u_n) \ge 1 - 2pe^{-c_1 u_n^2} \to 1, \tag{54}$$

since $\log p = O(n^\alpha)$. It follows from (27) and (28) in Condition 4, (48), (49) that

$$\begin{aligned} \|\mathbf{z}\|_\infty &\le (n\lambda_n)^{-1} \left[\left\| \boldsymbol{\xi}_{\mathfrak{M}_0^c} \right\|_\infty + \left\| \mathbf{X}_2^T \boldsymbol{\mu}\left(\widehat{\boldsymbol{\theta}}\right) - \mathbf{X}_2^T \boldsymbol{\mu}\left(\boldsymbol{\theta}_0\right) \right\|_\infty \right] \\ &= o(1) + (n\lambda_n)^{-1} \left[\left\| \mathbf{X}_2^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1 (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \right\|_\infty + \|\mathbf{w}\|_\infty \right] \\ &= o(1) + (n\lambda_n)^{-1} \left[O(n) \left\| \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \right\|_2 + O(n) \left\| \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \right\|_2^2 \right] \\ &= o(1) + O\left(\lambda_n^{-1}\sqrt{s/n}\tau \right) = o(1), \end{aligned}$$

which shows that inequality (8) holds for sufficiently large n. This concludes the proof.

8.8 Proof of Theorem 4

Clearly by Theorem 3, we only need to prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$. On the event H_n defined in the proof of Theorem 3, it has been shown that $\widehat{\boldsymbol{\beta}}_1 \in N_\tau \subset \mathcal{N}_0$ is a strict local maximizer of $\overline{Q}_n(\boldsymbol{\delta})$ and $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$. It follows easily that $\nabla \overline{Q}_n(\widehat{\boldsymbol{\beta}}_1) = \mathbf{0}$. In view of (52), we have

$$\nabla \overline{Q}_n(\boldsymbol{\delta}) = \nabla \overline{\ell}_n(\boldsymbol{\delta}) - \bar{p}_{\lambda_n}(\boldsymbol{\delta}).$$

We expand the first term $\nabla \overline{\ell}_n(\boldsymbol{\delta})$ around $\boldsymbol{\beta}_1$ to the first order componentwise. Then by (28) in Condition 4 and $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_2 = O_P(\sqrt{s/n})$, we have under the L_2 norm,

$$\begin{aligned} \mathbf{0} = \nabla \overline{Q}_n(\widehat{\boldsymbol{\beta}}_1) &= \nabla \overline{\ell}_n(\boldsymbol{\beta}_1) - n^{-1}\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ &\quad + O(1)\left\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right\|_2^2 \sqrt{s} - \bar{p}_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1) \\ &= n^{-1}\mathbf{X}_1^T\left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)\right] - n^{-1}\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ &\quad - \bar{p}_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1) + O_P(s^{3/2}n^{-1}). \end{aligned} \tag{55}$$

It follows from $\widehat{\boldsymbol{\beta}}_1 \in \mathcal{N}_0$, and $p'_{\lambda_n}(d_n) = o(s^{-1/2}n^{-1/2})$ in Condition 6 that

$$\left\|\bar{p}_{\lambda_n}(\widehat{\boldsymbol{\beta}}_1)\right\|_2 \leq \sqrt{s} p'_{\lambda_n}(d_n) = o_P(1/\sqrt{n}), \tag{56}$$

due to the monotonicity of $p'_{\lambda_n}(t)$. Combing (55) and (56) gives

$$\mathbf{X}_1^T \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \mathbf{X}_1^T\left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)\right] + o_P(\sqrt{n}),$$

since $s = o(n^{1/3})$. This along with the first part of (26) in Condition 4 entails

$$\mathbf{B}_n^{1/2}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) = \mathbf{B}_n^{-1/2}\mathbf{X}_1^T\left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)\right] + o_P(1), \tag{57}$$

where $\mathbf{B}_n = \mathbf{X}_1^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{X}_1$ and the small order term is understood under the L_2 norm.

We are now ready to show the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$. Let $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$, where $\mathbf{A}_n$ is a $q \times s$ matrix and $\mathbf{G}$ is a symmetric positive definite matrix. It follows from (57) that

$$\mathbf{A}_n\mathbf{B}_n^{1/2}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) = \mathbf{u}_n + o_P(1),$$

where $\mathbf{u}_n = \mathbf{A}_n\mathbf{B}_n^{-1/2}\mathbf{X}_1^T[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)]$. Thus by Slutsky's lemma, to show that

$$\mathbf{A}_n\mathbf{B}_n^{1/2}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \phi\mathbf{G}),$$

it suffices to prove $\mathbf{u}_n \xrightarrow{\mathscr{D}} N(\mathbf{0}, \phi\mathbf{G})$. For any unit vector $\mathbf{a} \in \mathbf{R}^q$, we consider the asymptotic distribution of the linear combination

$$v_n = \mathbf{a}^T\mathbf{u}_n = \mathbf{a}^T\mathbf{A}_n\mathbf{B}_n^{-1/2}\mathbf{X}_1^T\left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}_0)\right] = \sum_{i=1}^n \xi_i,$$

where $\xi_i = \mathbf{a}^T \mathbf{A}_n \mathbf{B}_n^{-1/2} \mathbf{z}_i [y_i - b'(\theta_{0,i})]$ and $\mathbf{X}_1 = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T$. Clearly ξ_i's are independent and have mean 0, and

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{var}(\xi_i) &= \mathbf{a}^T \mathbf{A}_n \mathbf{B}_n^{-1/2} \phi \left[\mathbf{X}_1^T \mathbf{\Sigma}\left(\boldsymbol{\theta}_0\right) \mathbf{X}_1\right] \mathbf{B}_n^{-1/2} \mathbf{A}_n^T \mathbf{a} \\ &= \phi \mathbf{a}^T \mathbf{A}_n \mathbf{A}_n^T \mathbf{a} \longrightarrow \phi \mathbf{a}^T \mathbf{G} \mathbf{a}\end{aligned}$$

as $n \to \infty$. By Condition 6 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sum_{i=1}^{n} E\left|\xi_i\right|^3 &= \sum_{i=1}^{n} \left|\mathbf{a}^T \mathbf{A}_n \mathbf{B}_n^{-1/2} \mathbf{z}_i\right|^3 E\left|y_i - b'\left(\theta_{0,i}\right)\right|^3 \\ &= O(1) \sum_{i=1}^{n} \left|\mathbf{a}^T \mathbf{A}_n \mathbf{B}_n^{-1/2} \mathbf{z}_i\right|^3 \\ &\leq O(1) \sum_{i=1}^{n} \left\|\mathbf{a}^T \mathbf{A}_n\right\|_2^3 \left\|\mathbf{B}_n^{-1/2} \mathbf{z}_i\right\|_2^3 \\ &= O(1) \sum_{i=1}^{n} \left(\mathbf{z}_i^T \mathbf{B}_n^{-1} \mathbf{z}_i\right)^{3/2} = o(1).\end{aligned}$$

Therefore an application of Lyapunov's theorem yields

$$\mathbf{a}^T \mathbf{u}_n = \sum_{i=1}^{n} \xi_i \xrightarrow{\mathscr{D}} N(0, \phi \mathbf{a}^T \mathbf{G} \mathbf{a}).$$

Since this asymptotic normality holds for any unit vector $\mathbf{a} \in \mathbf{R}^q$, we conclude that $\mathbf{u}_n \xrightarrow{\mathscr{D}} N(\mathbf{0}, \phi \mathbf{G})$, which completes the proof.

A Appendix

A.1 Three commonly used GLMs

In this section we give the formulas used in the ICA algorithm for three commonly used GLMs: linear regression model, logistic regression model, and Poisson regression model.

Linear regression. For this model, $b(\theta) = \frac{1}{2}\theta^2$, $\theta \in \mathbf{R}$ and $\phi = \sigma^2$. The penalized likelihood $Q_n(\boldsymbol{\beta})$ in (3) can be written as

$$Q_n(\boldsymbol{\beta}) = -\left\{(2n)^{-1} \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right)\right\}, \tag{58}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$. Thus maximizing $Q_n(\boldsymbol{\beta})$ becomes the penalized least squares problem. In Step 3 of ICA, we have $\widetilde{Q}_n(\beta_j; \widehat{\boldsymbol{\beta}}^{\lambda_k}, j) = Q_n(\boldsymbol{\beta})$, where the subvector of $\boldsymbol{\beta}$ with components in $\{1, \cdots, p\} \setminus \{j\}$ is identical to that of $\widehat{\boldsymbol{\beta}}^{\lambda_k}$.

Logistic regression. For this model, $b(\theta) = \log(1 + e^\theta)$, $\theta \in \mathbf{R}$ and $\phi = 1$. In Step 3 of ICA, by (30) we have

$$\widetilde{Q}_n(\beta_j; \widehat{\boldsymbol{\beta}}^{\lambda_k}, j) = \ell_n(\widehat{\boldsymbol{\beta}}^{\lambda_k}) + n^{-1}\left\{\mathbf{x}_j^T\left[\mathbf{y} - \boldsymbol{\mu}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}\right)\right]\right\}\left(\beta_j - \widehat{\beta}_j^{\lambda_k}\right) \tag{59}$$
$$- \frac{1}{2n}\left[\mathbf{x}_j^T\boldsymbol{\Sigma}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}\right)\mathbf{x}_j\right]\left(\beta_j - \widehat{\beta}_j^{\lambda_k}\right)^2 - \sum_{j=1}^{p} p_{\lambda_k}\left(|\beta_j|\right),$$

where the subvector of $\boldsymbol{\beta}$ with components in $\{1, \cdots, p\} \setminus \{j\}$ is identical to that of $\widehat{\boldsymbol{\beta}}^{\lambda_k}$, $\widehat{\boldsymbol{\beta}}^{\lambda_k} = (\widehat{\beta}_1^{\lambda_k}, \cdots, \widehat{\beta}_p^{\lambda_k})^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$, $\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}) = \left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, \cdots, \frac{e^{\theta_n}}{1+e^{\theta_n}}\right)^T$, and

$$\boldsymbol{\Sigma}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}) = \operatorname{diag}\left\{\frac{e^{\theta_1}}{(1+e^{\theta_1})^2}, \cdots, \frac{e^{\theta_n}}{(1+e^{\theta_n})^2}\right\}$$

with $(\theta_1, \cdots, \theta_n)^T = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}$.

Poisson regression. For this model, $b(\theta) = e^\theta$, $\theta \in \mathbf{R}$ and $\phi = 1$. In Step 3 of ICA, $\widetilde{Q}_n(\beta_j; \widehat{\boldsymbol{\beta}}^{\lambda_k}, j)$ has the same expression as in (59) with

$$\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}) = \left(e^{\theta_1}, \cdots, e^{\theta_n}\right)^T \quad \text{and} \quad \boldsymbol{\Sigma}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}) = \operatorname{diag}\left\{e^{\theta_1}, \cdots, e^{\theta_n}\right\},$$

where $(\theta_1, \cdots, \theta_n)^T = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\lambda_k}$.

A.2 SCAD penalized least squares solution

Consider the univariate SCAD penalized least squares problem

$$\min_{\beta \in \mathbf{R}}\left\{2^{-1}(z - \beta)^2 + \Lambda p_\lambda\left(|\beta|\right)\right\}, \tag{60}$$

where $z \in \mathbf{R}$, $\Lambda \in (0, \infty)$, and p_λ is the SCAD penalty given by (4). The solution when $\Lambda = 1$ was given by Fan (1997). We denote by $R(\beta)$ the objective function and $\widehat{\beta}(z)$ the minimizer of problem (60). Clearly $\widehat{\beta}(z)$ equals 0 or solves the gradient equation

$$g(\beta) \equiv \nabla_\beta\left\{2^{-1}(z - \beta)^2 + \Lambda p_\lambda\left(|\beta|\right)\right\} = \beta - z + \operatorname{sgn}(\beta)\Lambda p'_\lambda(|\beta|) = 0. \tag{61}$$

It is easy to show that $\widehat{\beta}(z) = \operatorname{sgn}(z)|\widehat{\beta}(z)|$ and $|\widehat{\beta}(z)| \le |z|$, i.e., $\widehat{\beta}(z)$ is between 0 and z. Let $z_0 = \operatorname{sgn}(z)(|z| - \Lambda\lambda)_+$.

1) If $|z| \le \lambda$, we can easily show that $\widehat{\beta}(z) = z_0$.

2) Let $\lambda < |z| \le a\lambda$. Note that g defined in (61) is piecewise linear between 0 and z, and $g(0) = \operatorname{sgn}(z)[-|z| + \Lambda\lambda]$, $g(\operatorname{sgn}(z)\lambda) = \operatorname{sgn}(z)[-|z| + (\Lambda+1)\lambda]$, $g(z) = \operatorname{sgn}(z)\Lambda p'_\lambda(|z|)]$. Thus it is easy to see that if $|z| \le (\Lambda + 1)\lambda$, we have $\widehat{\beta}(z) = z_0$, and if $|z| > (\Lambda + 1)\lambda$, we have

$$\widehat{\beta}(z) = \operatorname{sgn}(z)\frac{|z| - \Lambda\lambda(a-1)^{-1}a}{1 - (a-1)^{-1}\Lambda}.$$

3) Let $|z| > a\lambda$. The same argument as in 2) shows that when $|z| \le (\Lambda + 1)\lambda$, we have $\widehat{\beta}(z) = z_0$ if $R(z_0) \le R(z)$ and $\widehat{\beta}(z) = z$ otherwise. When $|z| > (\Lambda + 1)\lambda$, we have $\widehat{\beta}(z) = z$.

References

[1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 939–967.

[2] Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57**, 33–45.

[3] Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–1732.

[4] Breiman, L. (1995). Better subset regression using the non-negative garrote. *Technometrics* **37**, 373–384.

[5] Bunea, F., Tsybakov, A. and Wegkamp, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Elec. Jour. Statist.* **1**, 169–194.

[6] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313–2404.

[7] Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57**, 1413–1457.

[8] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407–499.

[9] Fan, J. (1997). Comments on "Wavelets in statistics: A review" by A. Antoniadis. *J. Italian Statist. Assoc.* **6**, 131–138.

[10] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–2637.

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.

[13] Fan, J. and Lv, J. (2009). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, to appear.

[14] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928–961.

[15] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the linear model. *J. Machine Learning Res.* **10**, 1829–1853.

[16] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109–148.

[17] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–332.

[18] Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics* **7**, 397–416.

[19] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

[20] Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

[21] Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

[22] Koltchinskii, V. (2009). Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.* **37**, 1332–1359.

[23] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Ann. Statist.*, to appear.

[24] Levina, E., Rothman, A. J. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Statist.* **2**, 245–263.

[25] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

[26] Lv, J. and Liu, J. S. (2008). New principles for model selection when models are possibly misspecified. *Manuscript.*

[27] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

[28] Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc. B* **70**, 53–71.

[29] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–1462.

[30] Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Konig, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006). Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology* **24**, 5070–5078.

[31] Rothman, A. J., Bickel, P. J., Levina, L. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

[33] van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645.

[34] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

[35] Vrahatis, M. N. (1989). A short proof and a generalization of Miranda's existence theorem. *Proceedings of American Mathematical Society* **107**, 701–703.

[36] Wainwright, W. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE Transactions on Information Theory*, to appear.

[37] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49–67.

[38] Zhang, C.-H. (2009). Penalized linear unbiased selection. *Ann. Statist.*, to appear.

[39] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–1594.

[40] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Res.* **7**, 2541–2567.

[41] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

[42] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509–1566.